DOCUMENT RESUME

ED 145 017 UD 017 318

AUTHOR Bessey, Barbara L.
TITLE Further Documentation of State ESEA Title I Reporting Models and Their Technical Assistance Requirements Phase I, Part One.
INSTITUTION RMC Research Corp., Bethesda, Md.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Office of Planning, Budgeting, and Evaluation.
REPORT NO RMC-UR-308
PUB DATE Aug 76
NOTE 64p.

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS *Educational Assessment; *Evaluation Methods; *Evaluation Needs; *Historical Reviews; *Measurement Goals; *Measurement Techniques
IDENTIFIERS *Elementary Secondary Education Act Title I; Local Education Agencies; State Education Agencies

ABSTRACT

This volume summarizes the first phase of the effort to document the Title I evaluation and reporting system and assess what technical assistance requirements would be necessary when the system is implemented. The report also includes a summary of all of the activities conducted thus far in the study, and a detailed description of the evaluation models and the common metric. The present project had three main objectives; (1) review the evaluation models from a methodological viewpoint, (2) visit all the state educational agencies (SEA) and a sample of local educational agencies (LEA) in each state to discuss the evaluation and reporting system, and (3) estimate the resources that would be required to implement the system. Other topics discussed include: (1) a detailed summary of the site visits, including the reactions of SEA and LEA personnel and estimates of system implementation requirements, (2) a summary of the Policy Advisory Panel, including the Panel's concerns and recommendations, (3) a discussion of the technical developments that occurred during Phase I, including a summary of the Technical Advisory Panel Meeting held at the outset of Phase I, and (4) a brief description of Phase II activities. (Author/AM)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED145017

UR

RMC Research Corporation
A Resource Management Corporation Subsidiary

RMC Report
UR 308

FURTHER DOCUMENTATION OF STATE ESEA TITLE I REPORTING MODELS AND THEIR TECHNICAL ASSISTANCE REQUIREMENTS PHASE I (PART ONE)

Barbara L. Bessey
Larry D. Rosen
Alice Chiang
G. Kasten Tallmadge

With the assistance of:

Alexander B. Cannara
Barbara M. Fagan
Richard T. Johnson
Sonia A. Johnson
Lawrence I. Weiner
Christine T. Wood

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CD017318

August 1976

Prepared for
U. S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Office of Education/Office of Planning, Budgeting, and Evaluation

2570 West El Camino Real, Mountain View, CA 94040

HOME OFFICE: 7910 Woodmont Avenue, Bethesda, Maryland 20014

Draft Report

RMC Report
UR 308

FURTHER DOCUMENTATION OF STATE ESEA TITLE I REPORTING MODELS, AND THEIR TECHNICAL ASSISTANCE REQUIREMENTS

PHASE I

(PART ONE)

Barbara L. Bessey
Larry D. Rosen
Alice Chiang
G. Kasten Tallmadge

With the assistance of:

Alexander B. Cannara
Barbara M. Fagan
Richard T. Johnson
Sonia A. Johnson
Lawrence I. Weiner
Christine T. Wood

August 1976

Prepared for

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education/Office of Planning, Budgeting, and Evaluation

The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

RMC Research Corporation
Mountain View, California

FOREWORD

This volume summarizes the first phase of the effort to document the Title I evaluation and reporting system and assess what technical assistance requirements would be necessary when the system is implemented. This report includes a summary of all of the activities conducted thus far in the study and a detailed description of the evaluation models and the common metric. To complete the study, it was necessary to meet with both a Policy and a Technical Advisory Panel, representatives from all 57 State Education Agencies, and personnel from several Local Education Agencies in each State. Phase II of this contract continues the documentation through investigations of several technical and methodological issues.

The RMC staff is extremely grateful to all those individuals who contributed to the success of this effort; in particular, to the U.S. Office of Education representatives and the Policy and Technical Advisory Panel participants whose critical comments were valuable in refining the system. We are especially indebted to Janice K. Anderson, our project monitor, whose technical support and unwaning enthusiasm for our efforts greatly facilitated our work. This effort, perhaps, should also be dedicated, at least in part, to all of the State Title I coordinators, evaluators, and the representatives from more than 400 districts, whose courtesy, patience, and genuine interest in the system sustained us throughout the year.

TABLE OF CONTENTS

		Page
FOREWORD		ii
I.	INTRODUCTION	
	Background	1
	Overview of the Current Study	4
II.	THE TITLE I EVALUATION AND REPORTING SYSTEM	
	General Description of the System	5
	The Evaluation Models	7
	The Normal Curve Equivalent, a Common Metric for Evaluation	13
	The Normal Curve Equivalent with Non-Normed Tests	16
	Forms and Instructions for the Evaluation and Reporting System	19
III.	SITE VISITS	
	Background	21
	Description of SEA and LEA Evaluation Practices	25
	Reactions of the SEAs and LEAs to the Evaluation and Reporting System	29
	Implementation Requirements	34
IV.	POLICY ADVISORY PANEL MEETING	
	Description of the Meeting	39
	Reactions of the Panel Members to the Evaluation and Reporting System	39
	Panel Recommendations	43
V.	TECHNICAL ADVISORY PANEL MEETING	
	Description of the Meeting	45
	Technical Discussions	45
VI.	TECHNICAL DEVELOPMENTS	51
VII.	PHASE II ACTIVITIES	53
REFERENCES		55

I. INTRODUCTION

This report documents an approach to large-scale educational evaluation that RMC Research Corporation has developed under contract with the U.S. Office of Education, Office of Planning, Budgeting, and Evaluation. Design of the evaluation and reporting system, which is intended for nationwide use in the ESEA Title I program, began in 1974. After an initial development period of one year, the current contract (OEC-0-300-75-0212) was undertaken to further document the system, to present it to all State Education Agencies (SEAs) and approximately three Local Education Agencies (LEAs) in each State, and to assess the nature and scope of the technical assistance that would be required for implementation of the system. The evaluation and reporting system differs substantially from efforts that utilize survey research techniques to conduct large-scale evaluations. Instead, the new system embodies an approach that focuses on meaningful evaluations at the local level, converts the outcomes to a common metric, and provides higher-level aggregations.

The remainder of this chapter briefly describes some of the historical background of Title I that led up to the RMC study entitled State ESEA Title I Reports: Review and Analysis of Past Reports, and Development of a Model Reporting System and Format (Gamel, Tallmadge, Wood, & Binkley, 1975). The present study is described in the following chapters.

Background

The Elementary and Secondary Education Act became law in 1965. Title I of this act authorizes financial assistance to LEAs that have concentrations of economically disadvantaged children. These funds are used to provide programs designed to meet the special needs of educationally disadvantaged children.

Title I is a categorical grant program for which Congress appropriates nearly two billion dollars annually. These funds are allocated to SEAs for distribution to their eligible and participating LEAs. The LEAs, in turn, bear primary responsibility for designing, implementing, and evaluating local Title I projects.

Title I legislation requires annual or periodic program evaluations at each administrative level. LEAs are required to evaluate local projects and submit annual evaluation reports to their SEAs. SEAs are required to summarize district data and submit periodic State reports to USOE. USOE draws upon the State reports, surveys conducted by the National Center for Educational Statistics, and special studies in making its annual report to Congress on the national impact of Title I.

Past efforts to provide a nationwide assessment of Title I have taken two forms: nationwide surveys and examinations of annual State Title I evaluation data. Past nationwide surveys have not provided a satisfactory assessment of Title I impact (USOE, 1968; USDHEW/OE, 1970;

Glass, 1970; USOE, 1971). The 1968 and 1969 National Surveys of Compensatory Education (USOE, 1968; Glass, 1970), for example, provided meaningful cognitive impact data on only a small proportion of the Title I population. These samples were too small to be representative of the Title I population and thus provided no meaningful picture of overall program impact.

Examinations of State Title I Evaluation Reports (Wargo, Tallmadge, Michaels, Lipe, & Morris, 1972; Gamel et al., 1975) indicated that these documents also failed to provide the information needed to assess nationwide program effectiveness. This failure is partly due to the SEAs' difficult task of summarizing LEA reports, which often differ widely in both the types of information reported and the presentation format.

Despite these difficulties, the State reports continued to be judged the most appropriate source of useful evaluation information. A decision was therefore made to attempt to upgrade the quality of the local evaluations and to standardize reporting practices. These feelings were reflected in the Education Amendments of 1974 (P.L. 93-380), which specified that:

> "Sec. 151. (a) The Commissioner shall provide for independent evaluations which describe and measure the impact of programs and projects assisted under this title. Such evaluations may be provided by contract or other arrangements, and all such evaluations shall be made by competent and independent persons, and shall include, whenever possible, opinions obtained from program or project participants about the strengths or weaknesses of such programs or projects.
>
> "(b) The Commissioner shall develop and publish standards for evaluation of program or project effectiveness in achieving the objectives of this title.
>
> "(c) The Commissioner shall, where appropriate, consult with State agencies in order to provide for jointly sponsored objective evaluation studies of programs and projects assisted under this title within a State.
>
> "(d) The Commissioner shall provide to State educational agencies, models for evaluations of all programs conducted under this title, for their use in carrying out their functions under section 143(a), which shall include uniform procedures and criteria to be utilized by local educational agencies, as well as by the State agency in the evaluation of such programs.
>
> "(e) The Commissioner shall provide such technical and other assistance as may be necessary to State educational agencies to enable them to assist local educational agencies in the development and application of a systematic evaluation of programs in accordance with the models developed by the Commissioner.

"(f) The models developed by the Commissioner shall specify objective criteria which shall be utilized in the evaluation of all programs and shall outline techniques (such as longitudinal studies of children involved in such programs) and methodology (such as the use of tests which yield comparable results) for producing data which are comparable on a statewide and nationwide basis."

Before an improved evaluation and reporting system could be designed, it was first necessary to determine the amount and types of information required by the Federal government to make educational and fiscal policy decisions. To make this assessment, RMC staff members met with educational policy makers in Congress and the Department of Health, Education, and Welfare (including the U.S. Office of Education and the Title I office).

RMC staff examined the current status of Title I evaluations by analyzing the State Title I Annual Evaluation Reports for the years 1969 to 1974. The reports were analyzed for content, adequacy and validity of data, and policy relevance. An attempt was made to aggregate common data and identify longitudinal trends.

Once these two initial steps had been completed, preliminary objectives for the new system were formulated. The feasibility of these initial objectives was then tested by querying State Title I Coordinators and representatives from the U.S. Office of Education. Their feedback resulted in substantial modifications to the initial plans for the Title I evaluation and reporting system.

The system that was eventually developed consisted of three evaluation models that were designed to collect cognitive impact information. Two of the models could be used with either standardized achievement tests (normed tests) or criterion-referenced tests (non-normed tests). One model required the use of nationally normed tests but has since been extended to accommodate non-normed tests. While the system allowed considerable flexibility in both model and test selection, all models produced project impact measures expressed in the same metric. This metric, the Normal Curve Equivalent, was created to provide an interpretable, sensitive, equal-interval scale of achievement that could be derived from national percentile ranks. A detailed description of the evaluation models and the common metric is presented in Chapter II of this report.

Although the evaluation and reporting system was developed with inputs from Federal, State, and local sources, additional feedback was solicited. A preliminary version of the evaluation and reporting system was taken to nine pilot states to obtain critical comments and constructive recommendations from SEA evaluators and Title I program personnel. Their comments were instrumental in effecting several changes to the system. The revised system was described in a report prepared for USOE (Gamel et al., 1975).

Overview of the Current Study

The present project had three main objectives: (1) review the evaluation models from a methodological viewpoint, (2) visit all SEAs and a sample of LEAs in each State to discuss the evaluation and reporting system, and (3) estimate the resources that would be required to implement the system.

The evaluation models were reviewed at the beginning of the project by a Technical Advisory Panel composed of experts selected to represent a wide range of educational evaluation areas. This panel discussed the strengths and limitations of each model, the aggregation of data at a national level, and related evaluation issues.

The site visits required extensive preparation. Materials were developed both to train the site visitors and to aid them in explaining the system to others. The staff members who had not participated in the developmental effort were given intensive training by those who had visited the pilot States the previous year. Following the training period, five two-person teams began the site visits. In addition to presenting the system to State and local personnel, the site visitors also collected information on perceived implementation problems and estimated additional costs that would stem from system implementation.

The Policy Advisory Panel, which worked closely with RMC throughout the system-development effort, was convened after about half of the site visits had been completed. A summary of all site visit activities was presented to the panel members who subsequently expressed their concerns and formulated recommendations to USOE and RMC regarding the evaluation and reporting system.

The submission of this report marks the end of Phase I of the project. Plans for Phase II are discussed in a later chapter.

The remainder of this report is devoted to the following topics:

- a description of the evaluation and reporting system, including its most recent modifications;
- a detailed summary of the site visits, including the reactions of SEA and LEA personnel and estimates of system implementation requirements;
- a summary of the Policy Advisory Panel, including the Panel's concerns and recommendations;
- a discussion of the technical developments that occurred during Phase I, including a summary of the Technical Advisory Panel Meeting held at the outset of Phase I; and
- a brief description of Phase II activities.

II. THE TITLE I EVALUATION AND REPORTING SYSTEM

The Title I evaluation and reporting system was developed to provide meaningful and comparable information on all Title I projects throughout the country and to permit aggregation of this information at the school, district, State, and Federal levels. An effort was also made to design a flexible system, one that would offer its users a choice of evaluation designs and tests. The system consists of three evaluation models. These models generate comparable achievement data expressed in a common metric.

Reporting forms were developed to facilitate data collection and analysis and to standardize data output. Instructions were also incorporated into the evaluation and reporting system to provide guidance in completing the forms. In this chapter, the system is described as it currently exists, including the modifications that were made as a result of panel recommendations and the site visits to the SEAs and LEAs. A general description of the system, including its input, output, and flow of information, is included, followed by a discussion of the evaluation models and the common metric. Information regarding the forms and instructions and how they have documented the system's changes over the year concludes the chapter.

General Description of the System

The Title I evaluation and reporting system provides a systematic method of collecting and reporting data from Title I projects. In this section, the system's data input, output, and flow of information are discussed.

Data input. The system requests a minimal set of data from all Title I projects. The data are not intended to assess compliance or to duplicate information reported for other purposes (e.g., detailed cost data). Rather, the information called for was identified by national educational policy makers as being relevant and useful for making legislative and administrative decisions regarding the Title I program.

Six types of information are collected: student participation, parent advisory councils (PACs), personnel, training, cost, and impact. The Student Participation form calls for an unduplicated count of the students served, broken down by grade and by public or non-public school attendance, and a duplicated count of the number of students receiving various types of services. The PAC form is primarily a checklist of school and district PAC activities. Under the Personnel heading, information is sought regarding the number and type of personnel employed by the Title I projects. Training information includes the type and amount of training given to project teachers and aides. Project cost data consist simply of total project funds, broken down by Title I and other supplementary sources. Impact data, which are collected only from those Title I projects serving the basic skills (verbal and numerical) areas, consist of achievement test scores processed to yield an index of cognitive growth.

Data output. The evaluation and reporting system generates comparable data from all Title I projects and aggregates these data at district and State levels. The ultimate products of the system are State Title I Annual Evaluation Reports containing specified information in a standard format. These reports provide State summaries of the six types of information described above.

A significant feature of the system is that, in addition to the aggregations, project-level data are maintained at each level of the system. This feature permits potential exemplary projects to be identified and examined more closely for possible packaging and dissemination. The concept of a project is critical to the structure of the reporting system. A project is defined as an instructional treatment with objectives, methods, materials, personnel, and activities that are uniform for all those it serves. A project may exist in one or more grades within a school, or in one or more schools within a district, or even in several districts. Many Title I projects are clearly defined by their application for funding. In cases where a single application results in the funding of several qualitatively different instructional treatments at different sites, the term project is reserved for the individual treatments and not for the funded composite.

The system also enables States to relate project gains to various project characteristics. These characteristics include hours per week of project exposure, total project hours, instructor-to-student ratio, and cost per student. Analysis of such relationships can assist decision makers in determining whether any of these variables have significant effects on project outcomes.

System information flow: The three-tier reporting system. The system can be characterized as three-tiered (building, LEA, SEA). Data originating at the building level are aggregated to produce LEA reports. Data contained in the LEA reports, in turn, are aggregated to generate the State reports. Project summary information is also produced for use at each level. For example, the LEA reports contain a summary of each project's gain and its characteristics. In this way, the reporting system can be useful, not only for USOE, but also for the participating schools, districts, and States.

Although reporting forms have been developed for use at the building, LEA, and SEA levels, the system is flexible enough to allow alternative methods of collecting and aggregating data. For example, some States may wish to aggregate and analyze data collected directly from the buildings, by-passing the LEA altogether. In other instances, the district may have building-level information readily available, making it unnecessary to use the building reports. In still other cases, an additional regional level of aggregation may be added to the data-collection process. Any data-collection procedure preferred by the schools, districts, or States is accommodated by the system, as long as the final State reporting forms are identical to those submitted by all SEAs.

The Evaluation Models

The evaluation models provide procedures for assessing the cognitive impact of Title I projects in the basic skills areas. These models are product rather than process oriented in that they yield summary data about project outcomes.

The basic evaluation scheme. The three evaluation models measure the cognitive impact of Title I projects using a common definition of the treatment effect. The gain attributed to the treatment is defined as the difference between the treatment group's performance on a post-treatment test and an estimate of what that performance on the same test would have been had the group not received the treatment. In other words, the project's cognitive impact is the observed posttest performance of the treatment group minus the expected no-treatment posttest performance. This relationship can be expressed as follows:

$$\text{Treatment Effect} = \left[\begin{array}{c}\text{observed}\\ \text{post-treatment}\\ \text{performance}\end{array} - \begin{array}{c}\text{expected}\\ \text{no-treatment}\\ \text{performance}\end{array}\right].$$

The observed post-treatment performance is simply the mean posttest score of the treatment group. The no-treatment expectation is an estimate derived from one of the evaluation models. The three evaluation models differ in the way they generate this no-treatment expectation.

The three evaluation models are: Model A, the Norm-Referenced Model; Model B, the Control Group Model; and Model C, the Special Regression Model. Each model has variations that enable it to be used with either normed (Model A1, Model B1, Model C1) or non-normed tests (Model A2, Model B2, Model C2).

The norm-referenced evaluation design generates a no-treatment expectation from the assumption that the treatment group will maintain its status relative to the national norm group from pretest to posttest without treatment. The control group model utilizes the posttest (or adjusted posttest) scores of a control group as the no-treatment expectation. The special regression design employs the mean posttest score predicted from a comparison group's regression line as the no-treatment expectation.

Each model's applicability, implementation requirements, and assumptions are discussed below. The common metric used for expressing the treatment effect is then described in detail. The procedures for expressing gains in the common metric are discussed after the more general description of the metric.

Model A: Norm-Referenced Model. It is expected that the norm-referenced model will be the most widely adopted of the models, since it most closely resembles the current evaluation strategy of pretesting and posttesting the treatment group and comparing the gain with a norm-based expectation. This model makes use of standard scores and percentiles rather than the more commonly employed grade-equivalent scores, because the measurement errors inherent in grade equivalents make them inappropriate for evaluation purposes. A detailed discussion of the deficiencies of the grade-equivalent score can be found in Ilmadge and Horst (1975).

Model A users may select any achievement test for which adequate normative data have been collected from representative samples drawn from either the national or local populations. The amount of gain attributed to Title impact is computed as follows: (a) determine the pretest percentile status of the treatment group from the norms tables; and (b) from the appropriate posttest norms table, obtain the expected no-treatment performance assumin that, without Title I treatment, the group would maintain its pretest stat with respect to the norm group. The treatment effect is the difference between the observed posttest mean score and this no-treatment expectation.

Several requirements must be followed when using Model Al:

1. The pretest should not be used to select project participants.

   Test scores can be thought of as having two components: a true component that reflects the student's actual ability level, and an error component that is unrelated to the true component. Although the two components are not related to one another, both contribute to the observed test scores and are thus correlated with them. This means that low observed scores will reflect a preponderance of negative measurement error and high observed scores will reflect a preponderance of positive error. The average error for the total group tested can be assumed to be zero.

   If children are selected for treatment because their test scores are below the average of the group tested, their observed scores will be systematically too low. Similarly, children selected for above-average performance will have observed scores that are too high. Immediate retesting of both extreme subgroups will result in test scores that are closer to the mean of the total group originally tested, since the average error for the retested subgroup should approximate zero on the second testing. This statistical phenomenon called "regression toward the mean" or simply "regression effect," artificially inflates the measurement of project gain since a biased (too-low) pretest mean is subtracted from an unbiased posttest mean. For this reason, the pretest should be administered after project participants have been selected.

2. Test at norming time.

   Because the norm group is used as a surrogate control group, comparisons with the normative data are valid only if the treatment students

are tested at approximately the same time of year as the norming sample. Since cognitive growth is a function of time during the school year, it is clearly inappropriate to compare the scores of two groups tested at different times.

The assumption is implicit in most norms tables that the norms are valid for a period of three or four months. This assumption is clearly not tenable. It is simply not the case that children learn nothing over this period of time and then show large gains overnight at the end of the period. If the empirical normative data are collected somewhere in the middle of the three or four month period, a "true" fiftieth-percentile child would appear to be below average at the beginning of the period and above average at the end. This situation produces artificial gains where pretesting is done before and posttesting is done after an empirical norming date.

The problem is further compounded by the test publishers' provision of interpolated or "created" norms for periods when no normative data are collected. Most publishers provide fall, winter, and spring norms, even though empirical data are collected at only one, or possibly two, of these times. These interpolated or projected norms are not adequate because they are generally based on the erroneous assumptions, at least for disadvantaged children, that cognitive growth is linear for the nine months of the school year with one additional month's gain occurring over the three summer months.

3. Use comparable pretest and posttest.

The same manufacturer's test, and the same level and form where possible, should be used for both pretest and posttest. Although conversion tables allow scores from one test level to be converted to scores on another test level, these tables contain some measurement error resulting from curve fitting, rounding, and successive transformations. Therefore, it is preferable to use just one level for both pretest and posttest. The appropriate test level, one in which most pupils score in its middle range, can be determined empirically. For Title I children, this often means choosing a lower level than that nominally appropriate for the grade in question.

The same form of the test should be used whenever possible because the test content for alternate forms is never exactly comparable. Levine and Angoff (1958) have shown that the practice effect of using the same form is negligible when testing is separated by several months.

4. Evaluate only project participants who have both pretest and posttest scores.

To obtain comparable pretest and posttest scores, it is essential that the scores be based on the same group of students. Otherwise,

the evaluation may be biased. For example, students may drop out of programs because they are too slow or because they have successfully mastered the material. In the first case, eliminating the low scores of slow learners from the posttest may raise the mean posttest score. In the latter case, the mean posttest score may be lowered by not including the scores of the successful students.

When non-normed tests are used with the norm-referenced model (Model A2), additional procedures must be followed to express the treatment outcomes in the common metric (see section entitled The Normal Curve Equivalent with Non-Normed Tests).

Model B: Control Group Model. The control group model is the classical experimental design in which the performances of treatment and control groups are compared. Ideally, individual students from a single population are randomly assigned to treatment and control conditions. However, since selection of Title I project participants is based on economic as well as educational need, the random assignment method is not usually possible. For Title I projects, this model can be used when two pre-existing groups are enough alike so that they can be considered random samples from a single population. The assumption is made that the two groups do not differ systematically in any educationally relevant way prior to the beginning of treatment. Under these circumstances, the mean posttest score of the control group constitutes the no-treatment expectation for the treatment group; that is, it provides a good estimate of how the treatment group would have scored had it not received the treatment.

Model B can also be used where small systematic differences are known or assumed to exist between the treatment and control groups. In such instances, a statistical adjustment is made to compensate for the initial inequality.

Model B is often difficult to implement in Title I projects because of the problems encountered in obtaining suitable control groups. The evaluator must ensure that the two groups are equivalent, or nearly so, on whatever dimensions are educationally relevant. These dimensions include sex, race, socioeconomic status, and so on. At the same time, the control group must not be constructed by matching children from clearly dissimilar groups on the basis of their pretest scores, since treatment and control group children will then be different distances from the means of their respective groups. Different proportions of negative error will be present in their scores and, upon retesting, they will show different amounts of statistical regression (Thorndike, 1942).

There are a few situations, however, where implementation of Model B is clearly feasible and desirable. For example, in projects that last half a school year or less, the pool of eligible students can be randomly assigned to sessions. Thus, students assigned to the next session can serve

as controls for those currently receiving Title I services. A second situation in which Model B may be appropriate exists in large, homogeneous districts where schools not eligible to receive Title I funds are quite similar to those that do. In such cases, the non-Title I schools may be used as controls for the Title I programs.

When a suitable control group is found, a pretest must be administered to both groups to ensure that they are similar. If the mean pretest scores differ, then the posttest scores must be adjusted to account for initial differences. Assuming these differences are due to sampling or other random sources of error, analysis of covariance, with the pretest score as the covariate, should be used to adjust the observed posttest scores. If the pretest differences are real and not due to random events (that is, if the two groups cannot be regarded as random samples from a single population), the principal-axis method of correction should be used as analysis of covariance would underadjust for initial between-group differences (Campbell & Erlebacher, 1970). The reader is referred to Tallmadge (1976) or Wainer (1976) for a detailed discussion of these correction methods.

The implementation of Model B is straightforward once a suitable control group is found. The treatment effect is simply the difference between the mean posttest scores of the control and treatment groups. Model B has two significant and practical advantages over Model A: (a) different tests or different levels of the same test may be used for pretest and posttest; and (b) testing times need not coincide with norming dates. When non-normed tests are used with the control group model (Model B2), additional procedures must be followed to express the treatment outcomes in the common metric (see section entitled The Normal Curve Equivalent with Non-Normed Tests).

Model C: Special Regression Model. In the special regression model, an originally intact group of students is divided into treatment and comparison groups based on a pretest cutoff score. The no-treatment expectation is derived by relating the regression of posttest scores on pretest scores for the treatment and the comparison groups. The treatment effect can be measured at two points: at the cutoff score (Regression Discontinuity Model) or at the mean of the treatment group's pretest score (Regression Projection Model). In the Regression Projection Model, developed by Tallmadge and Horst (1975), the comparison group's regression line is projected downward to the treatment group's mean pretest score, where it provides a predicted posttest score. This no-treatment expectation is compared with the treatment group's observed mean posttest score to obtain a measure of project impact. (See Figure 1 for a graphic representation of Model C.)

The Regression Discontinuity Model, developed by Campbell and Stanley (1963), measures the treatment effect at the cutoff score. Both models yield the same treatment effect if the two regression lines are parallel. In the Title I evaluation and reporting system, separate estimates of the treatment effect are calculated using each model. The estimate obtained from the Regression Projection Model is the one reported, however, as this

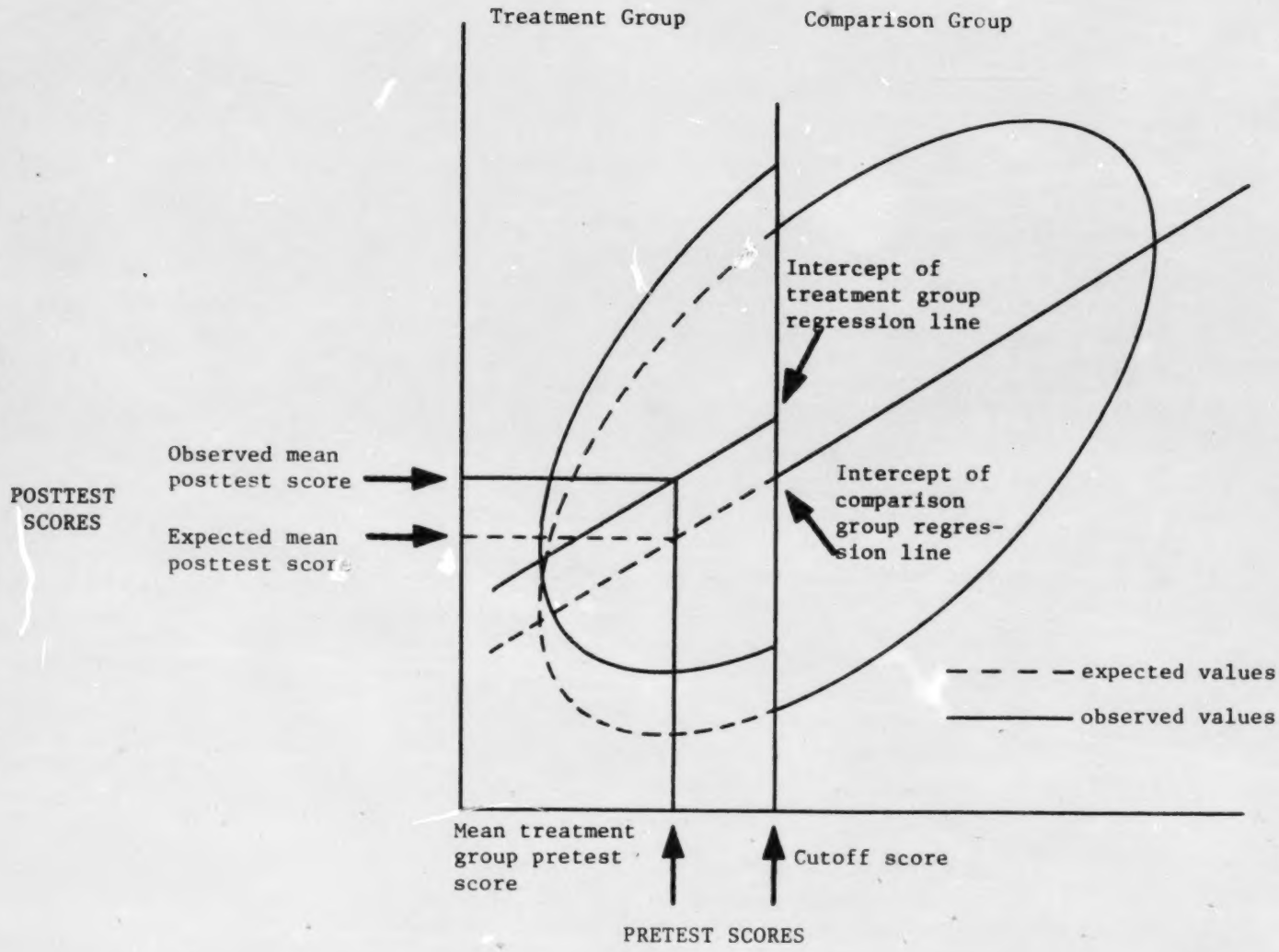

Figure 1. Graphic representation of Model C.

model is more sensitive in detecting a treatment impact that is inversely proportional to pretest scores (i.e., it is the lowest scoring students who make the biggest gains). The estimate obtained from the Regression Discontinuity Model is used to check for homogeneity and linearity of regression.

An underlying assumption of both regression models is a linear relationship between pretest and posttest scores. However, if either the floor or the ceiling of the pretest or the posttest scores is encountered, curvilinear regression will result. When Model C is applied under these conditions, a true treatment effect may be obscured or a false treatment effect may be created. To avoid these problems, it is essential that appropriate levels of the test be administered so that test floors or ceilings will not be encountered.

Since the special regression model necessitates pretesting and posttesting both treatment and comparison groups, it is most easily implemented in situations where data from school-wide or district-wide testing programs are available. Otherwise, the evaluator must test comparison group students as well as Title I students. To minimize the standard error of measurement, the sizes of the treatment and comparison groups should be relatively large.

There are few constraints concerning test selection. Unlike Model A, Model C does not require the same test for pretest and posttest or that the tests be administered at norming times. The only requirement is that the pretest and posttest instruments be highly correlated. A major difference between the norm-referenced model and the special regression model is that the pretest must be used as the selection device in the latter case. Procedures for using this model with non-normed tests are discussed in the section entitled The Normal Curve Equivalent with Non-Normed Tests.

The Normal Curve Equivalent, a Common Metric for Evaluation

The evaluation and reporting system is designed to produce impact data from different projects that can be compared and aggregated. The underlying assumption is that, for a given project, all models would yield identical estimates of the treatment effect. Thus, the outcomes from all three models are equivalent and can be aggregated.

Data-aggregation problems arise because different tests may be used in the system. Two questions are involved here: (a) Are the contents of different tests sufficiently similar to permit meaningful aggregation? and (b) Can scores from different tests be converted to a common unit of measurement? It is clear that each well constructed test has a unique content domain and its own content validity. There is, however, evidence that the most commonly used standardized achievement tests also have predictive or concurrent validity with one another (Loret, Seder, Bianchini, & Vale, 1974). Furthermore, other research has shown as high a correlation between a criterion-referenced test and a standardized test as between the standardized test and its own alternate form (Roudabush, 1975).

Data from tests with similar content or norming samples do not immediately lend themselves to aggregation and comparison. Existing test scores (raw scores, grade-equivalent scores, percentiles, stanines, or scaled standard scores) published by test manufacturers are inadequate for aggregating impact data because they are not comparable between tests. In other words, a standard score on one achievement test is not equal to the same standard score on a different test. The fact that some scores (e.g., percentiles and grade-equivalent scores) are not equal-interval scales further complicates data aggregation (Chiang & Rosen, unpublished). For these reasons a common metric was developed.

The national distribution of test scores provides the basis for the common metric--specifically, the national distributions of test scores at those grade levels for which normative data have been collected. The assumption is made that, after the score distributions are normalized, distance above or below the mean can be expressed in standard-deviation units that will essentially have the same meaning from test to test. This assumption is tenable if test contents are similar and if different test publishers have each tested truly representative national samples.

Normal Curve Equivalents (NCEs), the common metric for the system, are no more than normalized transformed standard (or z) scores. They have been given a mean of 50 and a standard deviation of 21.06 to produce a scale with the same range (1 to 99) and mean (50) as the percentile scale. NCEs can be generated directly from percentile ranks. Tables converting percentiles to NCEs and NCEs to percentiles are included with the system. Figure 2 depicts the relationship between NCEs, percentiles, stanines, and z-scores.

The treatment effects measured by all models in the system are expressed in the Normal Curve Equivalent metric. Using this scale, the results from different projects and different tests can be compared and aggregated. The following paragraphs briefly describe the methods employed by the models using normed tests to obtain the NCE treatment effect.

When norm-referenced tests are administered, the procedures for computing the NCE treatment effect for all three models are straightforward and similar: the NCE no-treatment expectation is subtracted from the observed mean NCE posttest score. In Model A, the NCE treatment effect is obtained by subtracting the treatment group's mean pretest NCE score from its mean posttest NCE score. In Model B, the NCE treatment effect is obtained by subtracting the control group's mean posttest NCE score from the treatment group's mean posttest NCE score. In Model C, the NCE treatment effect is obtained by subtracting the mean posttest NCE score predicted from the comparison group's regression line from the treatment group's mean posttest NCE score.

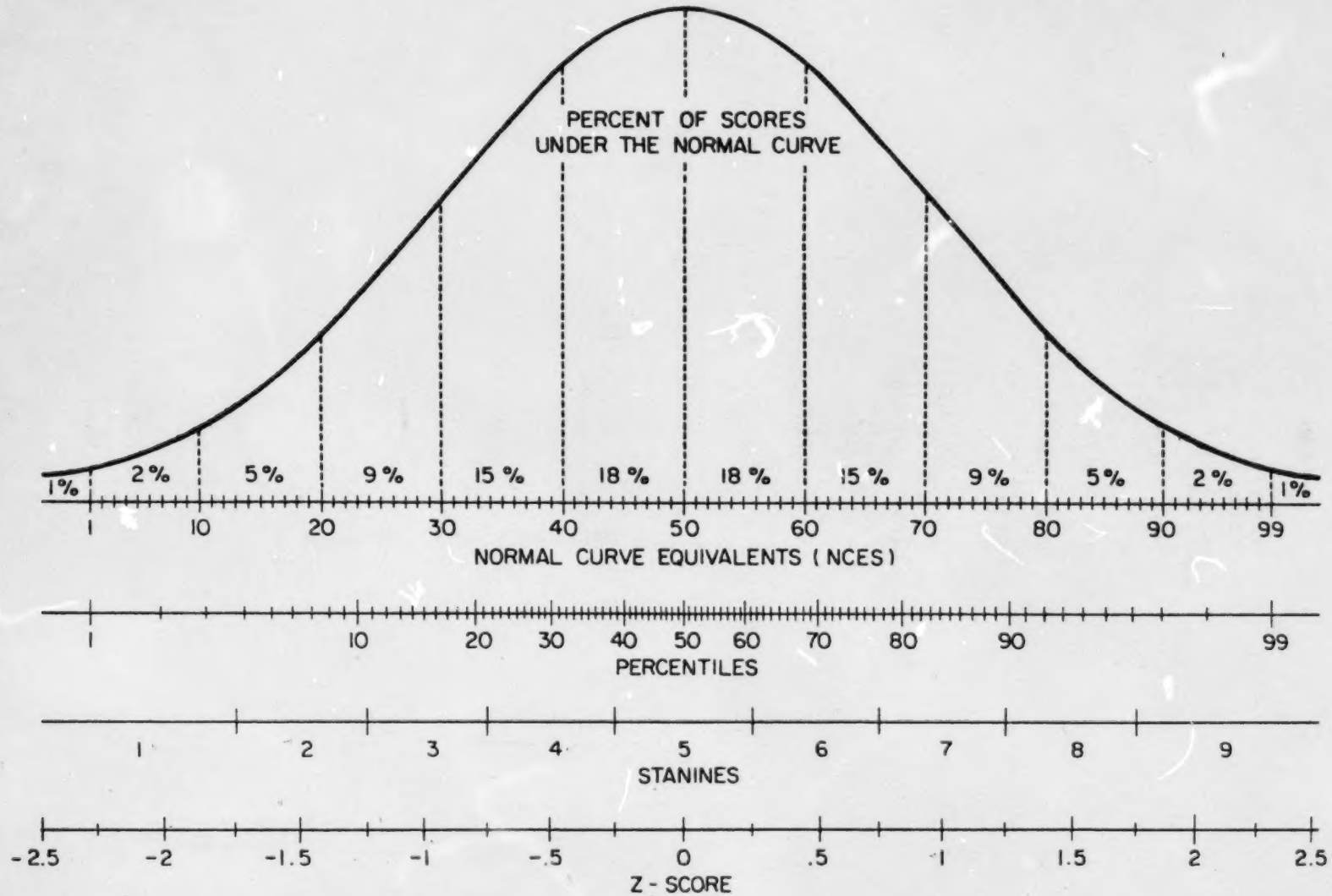

Figure 2. Relationship between Normal Curve Equivalents (NCEs), percentiles, stanines, and z-scores.

The mean NCE test score for the treatment, control, or comparison groups can be calculated either by finding the percentile equivalent of the group's mean standard score and converting the percentile rank to an NCE, or by converting the percentile score for each individual student to an NCE and computing the mean NCE. The method used depends upon the conversion tables provided by the test publisher; however, means can only be calculated from equal-interval test scores, such as standard scores or NCEs.

The methods used to obtain the NCE treatment effect when non-normed tests are used are described in the next section.

The Normal Curve Equivalent with Non-Normed Tests

An increasing number of evaluators are using criterion-referenced tests (CRTs) to measure Title I project impact. These tests are designed to measure individual performance with respect to a set of specified skills and not in relationship to the performance of a reference group. Since no norms are available for these tests, NCEs cannot be derived directly from the test scores. The system, however, was designed to accommodate both norm-referenced tests (NRTs) and criterion-referenced tests. This objective is achieved by requiring the administration of a nationally normed test to the treatment group in addition to the pretest and posttest CRTs. Methods for converting the treatment effect obtained from CRTs into the NCE scale are described below.

The treatment effect (TE) can be expressed as the observed mean posttest score ($\bar{Y}$) minus the expected mean posttest score ($\hat{\bar{Y}}$); that is,

$$TE = \bar{Y} - \hat{\bar{Y}} . \qquad (1)$$

The treatment effect can be expressed in standard deviation units by dividing the right side of Equation (1) by the national standard deviation (σ) and then converting to NCEs by multiplying the result by 21.06. That is,

$$TE_{NCE} = 21.06 \left(\frac{\bar{Y} - \hat{\bar{Y}}}{\sigma}\right). \qquad (2)$$

The values for $\bar{Y}$ and $\hat{\bar{Y}}$ can be obtained directly for each of the three evaluation models; for models using standardized tests, σ is available in the publisher's technical manual. However, since there are no empirical values of σ for non-normed tests, it must be estimated.

To do so, the following relationship between performances on the NRT and CRT is assumed:

$$\frac{\hat{\sigma}_{CRT}}{s_{CRT}} = \frac{\sigma_{NRT}}{s_{NRT}} \qquad (3)$$

where,

s_{CRT} = the standard deviation of the treatment group's posttest scores on the criterion-referenced test,

s_{NRT} = the standard deviation of the treatment group's scores on the norm-referenced test,

σ_{NRT} = the standard deviation of the national population on the norm-referenced test (found in the test publisher's technical manual),

$\hat{\sigma}_{CRT}$ = the estimated standard deviation of the national population on the criterion-referenced test.

This assumption maintains that the ratio between the sample standard deviation and the population standard deviation for a norm-referenced test is equal to the same ratio for a criterion-referenced test. This assumption can be verified empirically by collecting CRT scores from a nationwide representative sample and comparing the standard deviation with that predicted by Equation 3.

Solving for $\hat{\sigma}_{CRT}$, the estimated CRT population standard deviation, and substituting this value in Equation 2, the estimate of the NCE treatment effect when CRTs are used can be computed by the following equation:

$$TE_{NCE} = 21.06 \left[\left(\frac{\overline{Y} - \hat{\overline{Y}}}{s_{CRT}}\right) \times \left(\frac{s_{NRT}}{\sigma_{NRT}}\right)\right]. \qquad (4)$$

The administration of a standardized test to the treatment group is necessary to obtain the sample (s_{NRT}) and population standard deviations (σ_{NRT}) used in Equation 4. It is important that the NRT be administered on or very near the norming date. If the NRT is administered at any other time, σ_{NRT} must be derived by interpolating between successive normative data points.

When CRTs or non-normed tests are used with the control group model or the special regression model, the method described above for expressing the treatment effect in NCEs will yield results that are comparable to those estimated from the models using standardized tests. Because of the requirement that an NRT also be administered if CRTs are used for pretesting and posttesting, evaluators using the control group or the special regression models may find it convenient to administer a CRT as the pretest and use an NRT as the posttest. When non-normed tests are used with

the norm-referenced model, the procedures for computing the NCE treatment effect are further complicated by the need to derive the no-treatment expectation from normative data. These procedures are discussed below.

Model A with Non-Normed Tests. Model A2 has all of the assumptions associated with Model A1, plus others. Because of these requirements, Model A2 may appear less desirable than the other models. However, if it is impossible for an evaluator to construct control or comparison groups and if the selected test instrument lacks national norms, then Model A2 provides an acceptable evaluation design.

Model A2 requires that a CRT or other non-normed test be administered at pretest and posttest times to the treatment group children. A nationally normed test must also be administered at pretest time to the same group of children. Only the subtests of the NRT that measure the same skills as the CRT need to be given. However, all children tested within a grade must take the same form and level on the NRT. An abbreviated version of the equipercentile test score equating technique is used to "bridge the gap" between the observed median posttest score on the non-normed test, its normed test counterpart, and the corresponding national norm percentile value. Naturally, the larger the number of children tested, the more stable the NRT and CRT equating will be. An excellent presentation of this equating technique can be found in Angoff (1971).

For each test (pretest and posttest CRTs, and the NRT) the frequency and cumulative frequency of each raw score are computed. The assumption is made that NRT and pretest CRT raw scores can be equated if their corresponding cumulative frequencies are equal. However, it is not necessary to construct a table in which a raw score on one test is converted to a raw score on the other.

The equating of raw scores using their cumulative frequencies is necessary for calculating the observed posttest performance. The procedures for obtaining this observed posttest score are as follows: (a) determine the median raw score on the posttest CRT; (b) locate the cumulative frequency on the pretest CRT that corresponds to this median raw score value; (c) locate the NRT raw score that corresponds to the cumulative frequency just determined--this is the NRT raw score that corresponds to the median raw score on the CRT posttest; and (d) convert this NRT raw score to a percentile and then to an NCE. The no-treatment expectation is the percentile equivalent of the median raw score on the NRT, again converted to an NCE. The treatment effect is the difference between these two terms.

Model A2 makes the following assumptions in addition to those made in Model A1, and the credibility of the results of a Model A2 evaluation depends upon the degree to which they are justified:

- The norm-referenced test and the non-normed test must measure the same characteristics (i.e., both tests must measure reading ability or comprehension or vocabulary). The larger the correlation between the two tests, the more valid the resulting Model A2 evaluation will be.

- The relationship between NRT and CRT must be relatively constant between pretest and posttest. That is, the two tests must not interact with the treatment in different ways. This permits the equivalence tables developed between the two tests at pretest time to be applied to the CRT scores at posttest time.

- Both CRT and NRT must be reliable and valid tests.

Testing times are also important for a Model A2 evaluation. Both pretest and posttest administrations of the non-normed test must occur at times that correspond to the empirical normative data points for the normed test. To ensure a valid equating of NRT and CRT pretest scores, these tests should be administered as close together in time as possible. If the NRT is given before the CRT pretest, any practice effect that might occur as a result of taking the NRT will appear in both the CRT pretest and posttest scores and will not influence the gain. As with Model A1, selection of the project participants must be accomplished independently of both the normed and non-normed pretest scores. The use of pretest scores for selection invalidates the norm-referenced model.

Forms and Instructions for the Evaluation and Reporting System

The reporting forms that accompany the evaluation and reporting system are designed for use at the school building, the LEA, and the SEA. These forms standardize the collecting, analyzing, and reporting of data at all levels. Detailed instructions for completing the forms are also provided. These step-by-step implementation instructions can be followed easily by a technically unsophisticated user.

The forms and instructions, which are included in Part Two of this report, have undergone dramatic changes that reflect the system's development. The most important modification to the system is the change from building to project as the basic unit used for reporting impact information. This modification was adopted because it became apparent, after several site visits, that a report by building would not provide meaningful data in cases where several projects were implemented in a single building. Other efforts throughout the year were directed toward making the system more attractive to its users by increasing its flexibility. For example, Model A can now accommodate local norms. It can also be implemented using non-normed as well as normed tests. Model B has been extended to include appropriate statistical adjustments to compensate for nonequivalent control groups. These and other technical developments are discussed in Chapter VI.

Several features have been added to the forms and instructions to simplify data recording, analysis, and reporting:

- A glossary has been provided to clarify the meaning of terms used on the forms.
- Statistical notations on all forms have been replaced by words to facilitate understanding and reduce the "fear factor" among unsophisticated users.
- Appendices on system-relevant procedures, such as interpolating, developing local norms, and calculating correlation coefficients, have been included.
- An introduction to the evaluation and reporting system and an overview of each model now supplement the instructions.

III. SITE VISITS

Background

During Phase I, all of the States plus the District of Columbia, American Samoa, Guam, the Trust Territory of the Pacific, Puerto Rico, the Virgin Islands, and the Bureau of Indian Affairs, were visited by RMC site visit teams. The purposes of these visits were to:

- present and discuss the proposed Title I evaluation and reporting system with State and Local Education Agency personnel;
- collect information on potential problems and additional costs, which the SEA and LEA personnel felt they might incur as a result of implementing the system; and
- convince the SEA and LEA personnel of the desirability of adopting the proposed reporting system.

Part of the site visit effort was initially directed toward the goal of planning a field test of the evaluation and reporting system before initiation of a full-scale implementation. During the year, however, it became apparent that a field test of the size and scope envisioned at the outset of the project would not take place. Data-collection efforts were then redirected toward assessing the ramifications of system implementation. In particular, it was important for the subsequent planning of Title I efforts to ascertain what potential problems would arise from implementation and what additional costs would be incurred by SEAs and LEAs. Thus, although the goal of the project became modified, the objectives outlined for the site visits remained unchanged.

Training of the site visit teams. Five two-person site visit teams participated in several weeks of training. Training sessions were devoted to providing (a) historical background on the political and technical issues that relate to Title I, (b) background on evaluation in general, (c) knowledge on all aspects of the evaluation and reporting system, (d) opportunities to practice presenting the reporting system before critical audiences, and (e) in-depth exploration of any topics that the site visit teams felt merited further attention.

A syllabus of relevant reading materials was distributed to each site-visiting staff member during the training sessions. These materials provided more detailed information on the general history of Title I, its regulations and laws, the Education Amendments of 1974, test publishers' technical manuals, and other publications related to Title I evaluation.

The teams further prepared for the site visits by reading the Title I Annual Evaluation Reports of each State they would visit in order to familiarize themselves with the evaluation practices employed by the State. This advance preparation allowed the site-visiting teams to tailor their presentations to fit the specific situations encountered in a particular State.

The site visitors felt they had mastered the relevant materials by the end of the training sessions. The staff considered training to be an ongoing process throughout the year, however, and group sessions were held periodically to discuss questions, issues, and problems that were encountered on the site visits.

Site visit agenda. States were assigned to individual site visitors on the basis of economy and, whenever possible, personal preference. Other factors were also considered, such as rotating team membership and matching site visitors to States where enough was known in advance to suggest that some team members were better qualified than others to deal with the expected problems. The scheduling of the site visits was delegated to the individual teams. All 57 visits were successfully completed between mid-July 1975 and late May 1976. A complete list of the dates of all site visits is included as Table 1.

Although all of the site visits were similar in agenda and organization, flexibility on the part of the site visit team was encouraged. It became apparent that each State represented a unique set of circumstances in terms of the amount of time set aside for discussions with RMC personnel and the sophistication of the personnel in the area of evaluation. For the most part, however, visits to the SEAs lasted from three to five days, whereas one to two days were spent at each LEA.

At the SEA, the first day was devoted to the presentation of the models developed for Title I evaluation. A considerable period of time was allotted for discussions of the models and related evaluation issues. On occasion, two or more days were needed to present and discuss the models.

The reporting forms that accompany the evaluation models were presented on the second day. At this point in the visit, specific questions concerning the calculations of gain for each model were discussed. As on the first day, time was set aside for questions and further discussions.

On the third day, the SEA personnel worked through several specially prepared problem sets that entailed calculations at each level of the reporting system. By the end of this exercise, the SEA personnel had usually acquired a good understanding of the work requirements demanded by each level in the system. They were then in a position to indicate what potential problems might arise in the State were the system to be implemented, and whether implementation would require additional funding or personnel.

Additional days were frequently spent at the SEA performing optional activities, such as modifying the State's forms to conform to the requirements of the reporting system, analyzing State data using one of the evaluation models, discussing political or technical issues in greater detail,

TABLE 1

SCHEDULE OF SITE VISITS

STATE	DATE
Alabama	Mid-December
Alaska	Mid-September
Arizona	February/March
Arkansas	September/October
California	Mid-January
Colorado	Early September
Connecticut	Mid-October
Delaware	Late September
District of Columbia	October/November
Florida	Mid-November
Georgia	Early January
Hawaii	Early May
Idaho	Mid-September
Illinois	Late October
Indiana	Late May
Iowa	Mid-March
Kansas	Mid-February
Kentucky	Early May
Louisiana	Early November
Maine	July/August
Maryland	Early April
Massachusetts	Early September
Michigan	Late April
Minnesota	Early December
Mississippi	Mid-November
Missouri	Mid-May
Montana	Mid-March
Nebraska	Early January
Nevada	Mid-May
New Hampshire	Mid-August
New Jersey	Mid-April
New Mexico	Mid-August
New York	Early December
North Carolina	Early February
North Dakota	Mid-August
Ohio	Mid-October
Oklahoma	Mid-September
Oregon	Mid-November
Pennsylvania	March/April
Rhode Island	Mid-March
South Carolina	Late January
South Dakota	Late October
Tennessee	Early November
Texas	Mid-October
Utah	July/August
Vermont	Mid-November
Virginia	Mid-February
Washington	Late September
West Virginia	Late April
Wisconsin	April/May
Wyoming	Late August
American Samoa	Early March
Bureau of Indian Affairs	Mid-August
Guam	Mid-March
Puerto Rico	Mid-January
Trust Territory of the Pacific	Late March
Virgin Islands	Earl December

or presenting the system to additional audiences at the request of the SEA personnel.

The LEA visits were more varied than those to the SEAs. One reason was that State Title I Coordinators usually selected the districts and made the necessary arrangements with them in advance of the RMC visit. Although this procedure worked well for most of the site visits, there were occasions when State personnel were not entirely clear about the purpose of the site visit. As a result, meetings were scheduled with inappropriate district personnel, arrangements were made for irrelevant activities, or insufficient time was allotted for presentations. Often these problems were not recognized early enough to permit scheduling changes. Consequently, the site visit teams had to be flexible in adapting their presentations to whatever circumstances were encountered. Furthermore, meetings with the appropriate personnel in small districts often involved the Title I teachers, school principals, or other school staff. The fact that many of these individuals had to teach classes or attend to school concerns during the visit placed additional demands on the site visitors to adjust their presentations to fit local needs.

In spite of these difficulties, most of the LEA visits included a presentation of the models and the forms, and discussions of evaluation issues. Other activities, such as analyzing district data with one of the evaluation models, or working through the problems of aggregating data at different levels of the system with the LEA personnel, were included when appropriate.

During the course of the year, the site visit staff presented the evaluation and reporting system to personnel representing more than 400 districts, or an average of seven districts per State. These LEAs ranged in size from a one-building district with a Title I allocation of several thousand dollars to large metropolitan areas with allocations of several million dollars. As word of the RMC visits spread, States frequently requested large group meetings in addition to the three required LEA site visits. In fact, personnel from more than 90 LEAs in one State attended presentations of the reporting system. Requests by SEAs and LEAs for additional workshops were not uncommon. The Bureau of Indian Affairs (BIA), for example, requested that all ten, not just three, of its regions be visited by site visit teams. In lieu of these visits, approximately 20 BIA representatives gathered in the Mountain View office for a three-day "site visit" workshop on the proposed Title I evaluation and reporting system.

Site visit summaries. In addition to the presentations of the reporting system made to SEA and LEA personnel, the site visit teams also collected data that could be used to facilitate implementation of the system, plan future Title I studies, and design any field tests of the models. In particular, they were interested in how States and locals were evaluating existing Title I programs, how appropriate they felt the reporting system was to their needs, and what changes they would have to make to implement the system. The remainder of this chapter will summarize the staff's findings in the following areas:

- SEA and LEA evaluation practices encountered in the field;
- reactions of State and local personnel to the Title I evaluation and reporting system; and
- requirements for implementation of the proposed reporting system.

Description of SEA and LEA Evaluation Practices

Title I programs. A significant number of the Title I programs encountered were concerned only with the basic skills areas of reading or mathematics. Many of these programs were concentrated in the elementary grades even though Title I funds were available for programs from kindergarten through grade 12. The site visitors observed, however, that a number of States expressed an interest in performing evaluations of other Title I areas, such as early childhood, migrant, and neglected or delinquent programs. Inquiries were also made by some State and local personnel regarding procedures for evaluations in the affective domain.

Many States appeared to be cutting back on the number of support programs funded by Title I (e.g., health, counseling, and library) in favor of extending services in the basic skills areas. One of the reasons often given for this decision was that the SEA's interpretation of the Federal Title I regulations precluded the funding of many support programs. Generally, to receive auxiliary services, a child had to receive Title I services in at least one of the basic skills areas.

Evaluation: The selection of Title I students. The selection of Title I children was often based on single factors such as teacher judgment, a posttest score from the previous year, IQ, or a pretest score from the current year. Where test results were utilized in the selection process, children became eligible to receive Title I services by scoring anywhere from 1 to 20 months below grade level, the specific value often dependent upon the child's grade level.

In other instances, Title I selection was based on a combination of factors such as teacher judgment, test results, classroom performance, reading skills, mathematical skills, work habits, and attitudes. Many of the States' needs assessments made use of such combinations. Some even used sophisticated weighting procedures to arrive at a single needs assessment score for each child.

Because many States that employed a norm-referenced evaluation design were using a single set of test scores as both the selection instrument and as a pretest measure, the site visitors were often asked whether participant selection could be accomplished on the basis of teacher judgments. Resistance to this plan was often encountered from both SEA and LEA personnel. Some felt that selection based on teacher judgment, even in combination with other factors, was in clear violation of the Title I regulations. Others expressed concern that teachers would select children on inappropriate criteria, such as behavior and attitude, in order to have them removed from their classrooms.

In quite a few cases, the number of children who could actually participate in Title I programs was determined by a minimum per-pupil expenditure specified by the SEA. For many districts, this meant that the number of children that could be served was equal to their allocation divided by the minimum per-pupil expenditure. Selection of the actual Title I participants began with the neediest children and continued until the quota was filled.

Evaluation: Testing practices. A sizable number of LEAs employed a norm-referenced evaluation design in which a standardized achievement test was administered as a pretest in the fall and an alternate form of the same standardized test was administered as a posttest in the spring. Less common was spring-to-spring testing using standardized achievement tests. Fall-to-fall testing was even less frequently encountered, partly because fall testing would not allow enough time for the district to analyze its data or the SEA to write its report before the November 15 deadline set by the Office of Education for receipt of the State's Title I Annual Evaluation Report.

Very few States used criterion-referenced tests for evaluation purposes. Many, however, were interested in the possibility that CRTs could be used and often indicated that they felt they had to use standardized achievement tests because CRTs did not provide grade-equivalent scores. A sizable number of States that used standardized achievement tests to evaluate their programs employed CRTs extensively for diagnostic purposes, selecting program participants, or assessing progress within the Title I program.

Most States reported achievement test results in grade-equivalent scores. It appeared to the site visitors that State and local personnel did not really prefer grade-equivalent scores over any other derived test score. Grade-equivalent scores were used primarily to aggregate data, since no other common metric for comparing results from different tests was available. Some States also believed that the Title I regulations required the reporting of grade-equivalent scores.

SEAs often collected additional information to substantiate project gains. Percentiles, raw or standard scores, and stanines were reported by some States in addition to, or instead of, grade-equivalent scores. Several SEAs supplemented achievement data with anecdotes, testimonials, and narratives authored by Title I teachers, parents, or participants.

A few States had statewide testing programs to facilitate data collection at the building and district levels. Where this practice existed, however, test selection, and even testing times, were decided by the SEAs, leaving the LEAs without options to administer what they considered to be more appropriate tests at more appropriate times.

Although some States did not specify precisely what information they wanted from their LEAs, most provided some guidelines for LEAs regarding any or all of the following:

- selection of program participants: eligibility criteria (e.g., 6 months below grade level, test score below 33rd percentile, low score on State needs-assessment instrument);

- testing practices: what data to report (e.g., grade-equivalent scores, percentiles); what tests to use (e.g., any of the tests included in the Anchor Test Study, a CRT that was constructed for a particular State); or when to test (e.g., fall and spring, once-a-year testing in the spring); and

- cost: a minimum per-pupil expenditure; a minimum project allocation.

Evaluation: Data analysis. Two basic evaluation procedures were observed. In one case, the LEAs collected the data, conducted their own analyses, and sent the results to the SEA. In the other, the SEA collected raw data directly from the districts and performed all the analyses. The fact that raw data were sent directly to the SEA often indicated that the State had access to computer facilities. However, the site visitors encountered one State that performed all of its raw-data calculations by hand.

Some States performed extensive data analyses. In addition to reporting an overall project gain in reading or mathematics, the gain could also be broken down by numerous variables such as grade, project duration, IQ, sex, and so on. Multiple descriptions such as gain by project duration and sex, or gain by program area, grade, and IQ, were also observed.

Although most States collected data from every project, a few reported only the results from a sample of their districts. Many States expressed an interest in learning more about sampling procedures; they felt that a large number of LEAs coupled with an already overburdened staff made a total-population evaluation effort impractical.

In a significant number of States, SEA personnel stressed the importance of having districts specify behavioral objectives for projects and expectations for the project children. In fact, several SEAs indicated that they conducted district workshops that dealt exclusively with the topic of behavioral objectives. However, it was observed that most of these objectives were expressed in grade equivalents. Sample objectives might read as follows: 50% of the project children will make month-for-month gains in reading, or 75% of the project children will make more than 1.5 months gain per month in the mathematics project.

Although a majority of the States performed norm-referenced evaluations, most of them were not aware of the sound evaluation practices enumerated in the discussion of Model A. As a result, several problems were frequently observed:

- Project selection was based exclusively on the pretest: LEAs indicated that their teachers were wary that too much testing

was taking place. Furthermore, districts often wanted to exclude teacher judgment from the selection process. Therefore, to reduce the amount of testing involved or to ensure that selection was based on an objective measure, one test administration was used for both selecting and pretesting the treatment children.

- Tests were administered at inappropriate times: LEAs tested at convenient times with the test instruments of their choice. They felt that tests should be selected on the basis of quality and appropriate content, not only on norming times. Thus, tests with only one norm sample were used twice a year without employing the Anchor Test Study conversion tables.

- Grade-equivalent scores were used: States believed that the grade-equivalent scale was the only common metric available for aggregation across different tests at the present time. Many States and districts, although recognizing that grade-equivalent scores were inappropriate from a statistical point of view, felt that it was the only score parents, school boards, and legislators would understand.

- Difficulties were encountered with out-of-level testing: Title I children taking grade-level tests often performed at the chance level for that test. Because such test scores provided little or no useful evaluation information, a level of the test that was appropriate for the child's functional level was often administered. Many SEAs and LEAs, however, were unaware that out-of-level testing was either technically sound or politically acceptable. Some States, in fact, had official sanctions against out-of-level testing. A few States encountered a different problem. They discovered that the test of their choice did not provide expanded standard scores. This meant that out-of-level scores could not be converted to in-level norms--a situation that often made once-a-year testing impossible.

- Title I children were often highly mobile: A few districts reported that their project turnover during the year was greater than 100%. Although a turnover rate this extreme was rare, the high mobility rate was a problem for many districts. Some LEAs, in an attempt to compensate for this situation, initiated new procedures such as (a) installing a district-wide system for tracking Title I children if they moved within the district; (b) providing enough program flexibility so that a student who received Title I services in one school could immediately be placed in a Title I program in a new school without having to wait for a program opening; and (c) establishing similar Title I programs on a district-wide basis so that students coming from one school could fit into the same Title I program in another school with minimum disruption.

Evaluation: Capability for implementation. Should the evaluation and reporting system be implemented nationwide, the amount of technical assistance required by each State will depend, to a large extent, on the technical capabilities of its current Title I staff. Generally speaking, it is anticipated that those SEAs having personnel with strong technical backgrounds will be able to initiate system implementation with a minimum of outside assistance. Conversely, it is anticipated that those SEAs that lack technically capable personnel will need considerable technical assistance before implementation of the system can be successful.

Ease of system implementation may also depend upon access to computer facilities. When SEAs have access to data-processing facilities, raw data can be collected directly from the districts and analyzed by the SEA. Even when LEAs perform their own evaluations, SEAs often use their computers to examine more detailed relationships in the data, such as the amount of gain accounted for by several variables. Although some SEAs perform extensive analyses using only calculators, it seems that access to computer facilities permits the SEAs to spend more time on other aspects of Title I, such as program development. These SEAs may not be as likely to oppose reasonable changes to their current evaluation systems, since the amount of time spent changing existing computer programs would be minimal.

A surprising number of Title I evaluators have access, on a shared basis, to their SEA computers, but lengthy "turnaround" times make their use impractical. In these States and in those where no computer facilities were available, Title I personnel often felt overburdened by the existing evaluation requirements and were quite sensitive to any changes that might increase their workload.

Another variable, the use of an outside contractor to conduct the evaluation, may also be critical for implementaion of the system. Those SEAs and LEAs employing contractors seemed to enjoy a feeling of considerable freedom--it appeared that changes or modifications to existing evaluation efforts could be done with minimum effort since the contractor was hired to perform specified tasks.

Reactions of the SEAs and LEAs to the Evaluation and Reporting System

Although reactions to the evaluation and reporting system ranged from whole-hearted support, through mildly positive feelings or indifference, to vociferous disapproval, the majority of site visits produced positive reactions. A summary of these positive and negative reactions about particular models and the system in general is presented below.

General Comments. One of the feelings most often expressed about the site visits was appreciation for the effort to develop the system from the local level on up. SEA and LEA personnel were gratified that the models were developed with input from interested persons at all levels

of involvement in Title I. They liked the idea that modifications based on their comments would precede large-scale implementation.

Both SEA and LEA personnel were appreciative that this Federal effort, unlike several past efforts, did not require an audit of current evaluation practices. They were quite surprised to discover that the RMC site visitors came only to present and discuss the evaluation and reporting system and to solicit their comments and suggestions.

The fact that the system allowed marked flexibility in selecting instruments to assess program impact was viewed as an advantage by most SEA and LEA personnel. Many districts were pleased that they could finally use CRTs and still aggregate these data with those from districts using standardized tests.

Surprisingly, a substantial number of States and districts were agreeable to the replacement of grade-equivalent scores. These educators had already begun to understand that grade equivalents distorted their true program impact. The site visitors encountered many of these so-called inconsistencies in which a program would appear to be ineffective (e.g., making less than month-for-month gains) while the teacher was convinced that the students were learning.

Many SEAs and LEAs felt that the NCE common metric would provide a meaningful vehicle for comparing projects within and across districts. Only a few districts expressed reservations about being compared to other districts. Those that did express this view were often worried that, in the future, allocations might be tied to past performance or teachers might be evaluated on the amount of their program's impact.

Evidence regarding the positive impact of the site visits can be found in the fact that additional tasks were requested of the site visitors beyond the contract stipulations. Although RMC was required to visit only three LEAs in each State, SEA personnel often requested visits to additional LEAs or large workshops for many LEAs. These requests prompted a contract modification on October 21, 1975, that allowed RMC to hold group LEA meetings at central locations in these States. Following this modification, group meetings for substantially more than three LEAs were held in many States. In one State, a single meeting included representatives from 51 LEAs. Three similar workshops in that State were attended by an additional 48 LEAs for a total of 99 LEAs in the entire State.

Non-Title I personnel were also interested in additional assistance. Typically, following the site visitors' opening remarks, Title I personnel would suggest that many non-Title I personnel would also benefit from the discussions. Whenever possible, these educators were included in the Title I meetings. However, when these and other SEA staff members were unable to attend a particular meeting, the site visit team was often asked to hold special informal sessions.

In several States, the SEA evaluators and program planners did not wish to wait for a possible legislative mandate but planned immediate implementation of the evaluation and reporting system. Consequently, the site visitors were often asked to revise the existing evaluation data collection forms for the following school year.

Site visitors also converted prior years' evaluation data, expressed in a variety of ways, to Normal Curve Equivalents using one of the evaluation models. This exercise demonstrated to both SEA and LEA personnel that NCEs were easy to calculate and provided a better measure of project impact than other available metrics.

Not all reactions to the evaluation and reporting system were positive. Occasionally, the RMC site visit teams met with blanket disapproval. This disapproval was often directed at two sources: Federal intervention in Title I affairs, and the replacement of grade-equivalent scores with NCEs. A number of SEAs, especially those that believed they already had valid evaluation practices, felt some resentment for a Federal mandate that would force them to adopt a new, untested system. Because previous large-scale Federal evaluation efforts proved unsuccessful, many State and local personnel did not want to participate in another. The replacement of grade-equivalent scores with NCEs also turned out to be a controversial issue. A substantial number of SEAs argued that grade-equivalent scores should be replaced with NCEs for evaluation purposes but that grade-equivalent scores should be retained to communicate with parents, school boards, legislators, and so on. Because State and local personnel were very sensitive to both of these issues, heated discussions of these topics often took place.

A less frequent view was expressed by LEA personnel who disagreed with the intent behind the evaluation and reporting system. They indicated that no one in their LEA or on their school board really wanted to know if children received cognitive benefits from their Title I participation.

System comments. A primary purpose of the site visits was to assess whether the evaluation and reporting system was operationally feasible in various Title I situations. Accordingly, SEAs and LEAs were encouraged to react to the evaluation models and the reporting forms and instructions. Their critical comments are summarized below.

System comments: General. One of the greatest obstacles to overcome is the conflict between the proposed evaluation and reporting system and other Federal program policies. For example, while RMC site visitors were describing the hazards of grade-equivalent scores and suggesting their replacement with NCEs, some Federal program personnel were requesting grade equivalents and average monthly gains. Furthermore, many SEA personnel were somewhat annoyed that individual Federal programs had widely differing reporting requirements.

Surprisingly, there was less resentment than expected over the possible Federal mandate for adoption of the evaluation and reporting system.

SEAs and LEAs were generally willing to provide any data required by USOE as long as they were informed of specific requirements well in advance of the implementation date.

Other general comments included:

- a field test should be conducted before adoption of the system is mandated;
- the evaluation and reporting system should not emphasize only the basic skills;
- the term "project" should be carefully defined;
- sampling should be an alternative to reporting achievement data for all Title I students; and
- parent advisory council compliance information should not be included in the reporting system.

System comments: Model A. A substantial number of SEAs and LEAs preferred the norm-referenced model over the control group and special regression models. It is not too surprising that this model was favored since Model A is most similar to current evaluation practices. Model A also has the advantage that it does not require complex calculations or special technical training of personnel to implement it. In addition, Model A is also the least expensive model to implement because it does not need additional control or comparison groups or special computer programs to facilitate data analysis.

A few LEAs were interested in Model A2, the norm-referenced model using criterion-referenced tests. Generally, these districts received large allocations and had capable evaluation personnel.

Although Model A was generally favorably received, the recommended procedures for its implementation spawned considerable controversy. State and local personnel indicated that the requirement to test on norming dates might present problems. For example, many Title I programs want to pretest in the fall and posttest in the spring. Since only a few standardized achievement tests have collected normative data twice during the school year, test selection is limited. This means that many evaluators who test twice a year use tests normed only once during the school year.

Similar concerns were expressed over the recommendation to pretest and posttest with the same test level and form. This recommendation presents a problem for those programs in which students are tested only once during the school year. In these cases, each test is both a pretest and a posttest: the pretest for the subsequent year and the posttest for the previous year. Obviously, test levels must be changed periodically. When this occurs, additional testing may be required to compare the data from one test level with the data from the other level.

A number of State and local personnel were reluctant to compare their Title I students with the national norm. These educators felt that Title I students would not maintain their percentile rank without treatment and would fall farther behind the norm. Some who voiced this opinion were satisfied when it was explained that students who maintained their percentile status would fall farther behind the national norm in terms of grade-equivalent scores. Others continued to feel that a declining percentile expectation was more realistic. However, most felt that the no-treatment expectation issue is an empirical question that should be examined further.

System comments: Model B. Title I regulations currently specify that funds are to be used for Title I children only. In the control group model, as well as the special regression model, additional groups of children must be pretested and posttested. Much concern was expressed over the funding source for this testing. The consensus was that rules and regulations should be modified to allow Title I funds to be used for testing control or comparison groups. SEA and LEA personnel also felt that their evaluation budgets should reflect this additional cost.

A few districts were interested in Model B because they were considering a Title I evaluation using a nonequivalent control group design. The majority of LEAs, however, rejected the model outright because they felt it was a violation of both State and Federal guidelines.

System Comments: Model C. Model C was received with enthusiasm by many LEAs. Generally, these districts received large allocations, had capable evaluation personnel, and were interested in using criterion-referenced tests. In some cases, the LEAs had developed CRTs for their districts that they wished to pilot. These LEAs also preferred to select project participants on the basis of pretest scores and to avoid a no-treatment expectation based on maintaining a pretest national percentile.

Although criterion-referenced test users were interested in implementing this model, many districts expressed the reservation that the special regression model was statistically too complex to implement, even on a limited scale, without the aid of a computer program. They felt that Model C's applicability was therefore restricted to large districts with computing facilities.

System Comments: NCEs. Generally speaking, NCEs received positive responses. State and local personnel were pleased that criterion-referenced tests could now be used to evaluate Title I programs and that the resulting data could be aggregated with data from standardized tests. Two important reservations, however, were expressed about their interpretability. First, although grade-equivalent scores are fraught with systematic measurement errors, they are "meaningful" to teachers, parents, school boards, and Federal decision makers. NCEs are not yet meaningful to these groups. Many personnel felt that referencing NCEs to percentile

ranks would make them more meaningful; an equal number, however, felt that even percentiles would not be understandable to all. Others thought NCEs and percentiles would be confused.

A second interpretability issue concerned the size of an exemplary gain. A previous report (Horst, Tallmadge, & Wood, 1975) established a gain of one-third of a standard deviation, or approximately 7 NCEs, as educationally significant. However, many felt that this standard should be empirically tested.

Implementation Requirements

By the end of a site visit, not all SEA personnel could indicate to their satisfaction what problems they might face or what additional costs they might incur from system implementation. This situation, coupled with the fact that potential implementation problems are unknown, suggests that the opinions expressed in this section should be considered tentative and a first approximation only. Because each State and Local Education Agency is unique in terms of capabilities, resources, and other variables, generalizations are difficult. Therefore, this section will approach the implementation requirements from a qualitative and descriptive, rather than a quantitative, point of view and will include discussions on:

- factors influencing ease of implementation;
- changes in evaluation practices required for implementation; and
- potential additional costs.

Factors influencing ease of implementation. Model selection is an important factor as far as implementation is concerned. Feedback from the SEA and LEA personnel indicated that not much additional cost or personnel would be required for immediate implementation of Model A. However, it was felt that Models B or C would necessitate technical assistance, computer facilities, and additional testing funds.

LEA population size may prove to be important in determining implementation ease. However, many marked differences in evaluation practices were found among LEAs. Large LEAs, generally located in metropolitan areas, often had computer facilities, departmentalized evaluation units, technically capable personnel, and extensive communications with the SEA. All of these factors are conducive to successful implementation of a model. It appeared to be the case, however, that these large LEAs often had a bigger stake in current practices and were reluctant to abandon them.

On the other hand, small districts lacked many evaluation aids, but because they tended to follow the SEA's directives without question, they expressed little concern over adopting new practices. Although the personnel in small districts might not comprehend all the statistical underpinnings of the evaluation system, they would contribute to a successful implementation by providing full cooperation in data-collection

efforts. This is only a general observation. Several large districts visited had neither computer facilities nor an evaluation staff. Similarly, some small districts were quite technically sophisticated and had computer facilities.

An important consideration might also be the availability of data-processing facilities, a factor that may be correlated with LEA size. The notion that computer access allows greater flexibility for change in evaluation practices has already been discussed in an earlier section of this chapter (Description of SEA and LEA Evaluation Practices). Compute services also afford evaluators an opportunity to analyze their data using more sophisticated statistical techniques, such as those found in Models B or C. In addition to reducing calculation errors, computer processing also makes it possible to analyze test data using more than one model with minimum effort.

Where SEA or LEA evaluation practices are similar to those proposed by the reporting system, implementation of the models should not prove difficult. For example, if SEAs are currently collecting individual pretest and posttest percentile scores, then adding a step to convert percentiles to NCEs is quite simple.

One influential factor that should not be underestimated is the relationship between the SEA and its LEAs. This relationship appears to be affected in part by the number of LEAs in the State. In cases where the SEA visits each LEA during the course of the year, feelings of mutual support and cooperation were generally observed. However, in other States, dissemination of information was left to periodic workshops, and personal contact between State and districts was at a minimum. In many such districts, the personnel indicated that they did not know whom to call with evaluation problems, a situation that is clearly not conducive to uneventful implementation of a new evaluation system.

Changes in evaluation practices required for implementation. Before existing practices can be changed, a large-scale dissemination effort must be directed at potential system users to inform them of the evaluation models and their requirements. In addition to educating LEAs and Title I school personnel, parents, school boards, and other community groups must also be informed. This effort may prove difficult for States that lack effective communication links with their districts.

Model A, although computationally the easiest to use and clearly the model most often preferred by SEAs and LEAs, has several requirements that must be closely followed. These requirements, however, are being violated by current practices. The problems and approaches to their solutions include:

- The test used to select project participants cannot also be used as a pretest: alternate selection methods offered to SEAs and LEAs for future consideration were previous test scores, teacher judgments, diagnostics, and so on. Although

these suggestions were occasionally regarded as unsound, illegal, or both, they offer the only alternative to additional testing.

- Tests must be administered at times that correspond to empirical normative data points: as indicated earlier, test selection is limited for those evaluators who wish to test twice a year. In an effort to improve the existing situation, RMC personnel initiated discussions on this issue with several test publishers, some of whom have already modified their data-collection procedures to provide more usable norming dates.

- Functional-level rather than grade-level tests should be administered: as mentioned above, functional-level testing, which may often involve out-of-level tests, is a way of obtaining meaningful evaluation data for Title I children who may score at the chance level on the test nominally appropriate for their grade. However, expanded standard scores, which are essential for converting out-of-level scores to in-level scores, are not provided by all test publishers. In an attempt to remedy this situation, RMC personnel have initiated discussions on this issue with several test publishers to make them aware of the problem. An additional problem with functional-level testing is that more levels of a test must be purchased, making careful accounting and test management essential.

Fewer changes in actual testing practices would be required for implementation of Models B or C. Part of the reason for this is that some of the requirements that are necessary for a successful implementation of the norm-referenced model, such as testing times and selection procedures, are not important to the control group or the special regression models. However, additional tests would have to be purchased and administered to the no-treatment children, and the resulting additional data collected and analyzed.

It is understood that implementation may also necessitate modifications to existing data-collection forms. In particular, grade-equivalent scores must be replaced by standard scores, percentiles, or NCEs. A more difficult change that SEAs and LEAs will have to make is the reformulation of program objectives that are couched in grade-equivalent gains. More empirical tests of other methods, such as those based on percentiles, NCEs, or specific program goals, are needed to determine a reasonable alternative to the present system. Since NCEs and percentile ranks are less intuitively meaningful than grade-equivalent scores, further familiarization will be required.

Potential additional costs. Five sources of potential additional cost were identified: workshops, personnel, modifications to current evaluation systems, computer programs, and testing. Regardless of how much additional cost would be incurred, all SEA and LEA personnel felt that their overall Title I evaluation budgets should reflect these increases.

Workshops seemed to be the most popular method used by State and district personnel to disseminate information and to answer questions concerning Title I policies. Almost everyone agreed that additional workshops on the reporting system would be needed. Some SEAs felt that they would benefit from a "refresher" course on the system, while others felt that State-level workshops should be directed toward accountants, program personnel, or others who were not included in the RMC site visit discussions. SEAs were particularly interested in obtaining informative materials about the system to use in their district workshops.

Most LEAs also felt the need for workshops. Although some felt that only the models selected by the SEA (or LEA) should be presented at the district workshops, others felt all models should be included to allow flexibility in model choice. State personnel, in particular, wanted LEAs to receive enough technical assistance, since the system was designed so that almost all of the calculations were performed at the district level. SEA and LEA personnel felt quite strongly that technical assistance on a continuous basis should be available for successful nationwide implementation of the evaluation and reporting system.

LEAs felt that specialized workshops would be required to educate and train teachers, principals, and other building-level personnel to use the reporting system. Workshops, pamphlets, or other printed materials were felt to be appropriate for communicating the system to parents, school boards, legislators, and others.

Since many State and local personnel appeared to be overburdened with their current Title I activities, it was not surprising that a need for additional personnel was frequently expressed. Many LEAs, regardless of staff size, felt they would need more clerical help. In smaller districts, where one person was responsible for both evaluation and program development, a need was expressed for an additional professional person to share administrative responsibilities. SEAs often indicated a need for additional personnel to conduct district workshops, visit LEAs, analyze data, and so on.

One issue concerning staffing that arose was unexpected. Since a concern was expressed whether proper test administration procedures were carried out at pretest and posttest times, SEAs and LEAs felt that additional building-level personnel were needed to collect the test data. Recognizing that this request might be too expensive or impractical, they suggested that some type of incentive or bonus be given to the current testing staff at the building level to encourage accurate data collection.

Not all SEAs and LEAs requested additional personnel; many felt their current staffing arrangements were adequate. However, some felt that funds would be necessary to modify their existing evaluation programs. These modifications would include updating evaluation or application forms, maintaining appropriate record-keeping procedures, informing personnel of the new system, and updating computer programs. These two latter points are expanded below.

For SEAs and LEAs with access to computer facilities, additional funds were requested for such items as updating current data-analysis programs, developing new programs, purchasing machine time, and maintaining an adequate programming staff. State personnel also indicated that more funds would be necessary to enlarge current data bases to include data for control (Model B) or comparison group (Model C) children. Others felt that additional funds would be needed for scoring the no-treatment children's tests on their computers.

For each model, additional funds may be required for testing:

- Model A: A separate test may have to be purchased, administered, and scored to determine treatment group selection; more test levels may have to be purchased to allow functional-level testing.

- Model B: Pretests and posttests must be purchased, administered to control group children, and scored; an additional norm-referenced test must be administered and scored if criterion-referenced tests are used; existing computer data bases must be enlarged to include test scores from control group children.

- Model C: Posttests must be administered to comparison group children and scored; an additional norm-referenced test must be administered and scored if criterion-referenced tests are used; existing computer data bases must be enlarged to include test scores from comparison group children.

Although each of these areas was singled out as a potential source of additional cost, conclusions should be drawn only tentatively. Once implementation is underway, some costs may prove to be negligible, while others may be substantial.

IV. POLICY ADVISORY PANEL MEETING

The Policy Advisory Panel met with the RMC project staff four times during the previous contract. Throughout these meetings, the panel was particularly helpful in pointing out the positive aspects of the system as well as specific problem areas that needed additional work. The panel was convened during the present project. A description of the meeting is presented below, along with the panel's reactions and recommendations.

Description of the Meeting

A Policy Advisory Panel meeting was held February 24, 25, and 26, 1976, in Palo Alto, California. A list of the participants and the agencies they represented is found in Table 2. Several objectives were established for the conference. In particular, the RMC staff planned to:

- bring the panel members up to date on the recent developments of the evaluation and reporting system;
- present a summary of what was learned during the course of the site visits to the SEAs and LEAs, including current evaluation practices, reactions to the reporting system, and judgments regarding requirements for implementation of the system;
- discuss the revised forms and instructions; and
- elicit reactions from the panel members on several issues related to the evaluation and reporting system, including recommendations for implementation.

Reactions of the Panel Members to the Evaluation and Reporting System

The panel participants felt that more than the allotted 2-1/2 days was needed for adequate coverage of the RMC presentations and the subsequent discussions. In particular, the panel members felt that they needed more time to formulate their recommendations.

Each participant brought different concerns to the discussions of the models and the reporting system. Since one of the goals of the meeting was to elicit constructive comments about the system and reactions to related evaluation issues, the remainder of this chapter will be devoted to presentations of the panel's comments. For the sake of brevity, only those opinions shared by several participants will be discussed. Despite the fact that many negative comments were voiced, the overall reaction of the panel was quite positive. The criticisms, however, appeared to reflect a genuine effort on the part of the panel members to improve the system.

Positive comments. Generally, the panel continued to react favorably toward the evaluation and reporting system. Several reactions expressed

Table 2

POLICY ADVISORY PANEL MEETING
February 24, 25, 26, 1976

POLICY ADVISORY PANEL

Gary Brummitt	Missouri SEA
Nancy Burton (1)	National Assessment of Educational Progress
Richard Faunce	Minneapolis LEA
Constance Gomes	Title I Parent (Rhode Island)
Weldon Idol	Winston-Salem LEA
Rosella Lipson (2)	National Advisory Council of Disadvantaged Children
Rose Mary Loving	Title I Parent (Louisiana)
Vincent Madden (3)	Council of Chief State School Officers, Committee on Evaluation Information Systems (CEIS)
Wallace Roby	Connecticut SEA
Paul Rost	Albuquerque LEA
Daniel Schooley (1)	Michigan SEA

U.S. OFFICE OF EDUCATION

Janice Anderson	Office of Planning, Budgeting, and Evaluation
Ronald Fishbein	Office of Planning, Budgeting, and Evaluation
Gerry Hendrickson	Office of the Assistant Secretary for Planning and Evaluation
William Lobosco	Division of Education for the Disadvantaged
George Mayeske	Office of Planning, Budgeting, and Evaluation
Carl Wisler	Office of Planning, Budgeting, and Evaluation

RMC PROJECT STAFF

Barbara Bessey	Sonia Johnson
Alexander Cannara	Oscar Roberts
Alice Chiang	Larry Rosen
Barbara Fagan	G. Kasten Tallmadge
Richard Johnson	Christine Wood

(1) Added after original Policy Advisory Panel was formed.
(2) Attended for Allen Davitt.
(3) Represented the California SEA in previous panel meetings; at this meeting, in addition to representing the California SEA, he also replaced Andrew Nutt as the representative from CEIS.

by the participants in previous meetings were reiterated in the current discussions. In particular, they liked the fact that alternative models offered flexibility; that selection of any particular model could be made by building, LEA, or SEA personnel; that the system generated comparable data nationwide regardless of the test instrument selected; and that the Normal Curve Equivalent was a viable alternative to the grade-equivalent score.

Most panel members felt that the RMC monograph entitled A Practical Guide to Measuring Project Impact on Student Achievement (Horst, Tallmadge, & Wood, 1975), which had been published since the last panel meeting, was a valuable and important document. Moreover, the panelists who had already been visited by an RMC site visit team, agreed that the site visit had been worthwhile--as a "refresher" course for them, as well as a good introduction for their colleagues.

Some SEA and LEA evaluators on the panel, who have already begun implementation of a model, expressed the feeling that the system was easier to use and involved less effort than previous evaluation plans. However, most panelists did agree that the development of a computer program by RMC was essential for satisfactory large-scale implementation of the models.

Although most participants felt that the revised edition of the forms, which included the idea of project-level reporting, was superior to previous versions, they also indicated that more effort was needed to clarify terms that would make both the forms and the instructions easier to follow.

General comments. The discussions by the panel participants centered around the issues generated by the technical aspects of the system and the problems of disseminating the new system to its potential users.

The system's new technical developments were carefully scrutinized. Panel members felt that overlapping membership between the Policy and Technical Advisory Panels was needed so that the problems associated with delineating good field evaluation and rigorous statistical techniques could be shared.

Although the panelists recognized that the evaluation and reporting system is still being refined, they expressed an urgent need for a guidebook specifying when, where, and how each model may be applied. RMC staff members are currently writing such a monograph, which will serve as a user's guide to the system. Furthermore, the panel felt that refinement of the reporting system and, in particular, the models and their underlying technical assumptions, should be a continuous process.

All SEA and LEA panel representatives were particularly concerned with the logistics of USOE's technical assistance plan. All felt that some technical assistance would be necessary for the Local as well as the State Education Agencies. However, disagreement arose as to whether this assistance should come directly from the U.S. Office of Education to the SEAs, or from an intermediate agency inserted between USOE and the

States. To facilitate implementation, many panelists felt that the technical assistance directed toward the SEAs should include the larger districts as well, since many of these districts have both technical expertise and access to computer facilities.

Although no one questioned the importance of parental involvement, several participants did believe that compliance items, namely those in the current parent advisory council forms, should not be included with impact data collection. Although the intent of these forms was not to determine compliance, but rather to assess parental involvement, the fact remained that a district that failed to check a particular item would be in violation of the Title I regulations.

The issue of a field test generated considerable discussion. Part of RMC's efforts in Phase I were originally directed toward planning a field test. During the field test, all models would have been implemented on a pilot basis so that potential problems with the models or the forms and instructions could have been solved before full-scale implementation was undertaken. However, for a number of reasons, the U.S. Office of Education decided to proceed with system implementation plans without the large-scale field test. Several panel members felt quite strongly that some field test of the system should occur prior to nationwide implementation.

Almost all panel members expressed some concern as to how the system could best be communicated to the necessary audiences such as Congress, school boards, school principals, teachers, and parents. The feeling was expressed that, since the concepts behind NCEs might be too technical for the general public, percentiles should be used for communication. However, others thought that NCEs could be communicated without too much difficulty, as long as the descriptions were in language the public could understand. Thus, it was suggested that RMC write a series of articles on the reporting system, and NCEs in particular, that would appear in journals catering to various educational and community groups, such as parents, principals, and teachers. Since most of the panelists supported parental involvement in reviewing evaluation procedures, it was felt that some knowledge of the system by parents was essential.

System comments. The panel indicated that one of the major liabilities of the evaluation and reporting system was that only one aspect of Title I programs, namely cognitive impact in the basic skills areas, was evaluated. Panel members were concerned that other programs, such as those in the affective domain or in early childhood education, were not considered. However, the U.S. Office of Education did indicate that development of evaluation models for the other Title I areas was currently underway or planned for future efforts.

A concern was also expressed that no consideration was given to any long-term maintenance of the system. The considerable efforts on the part

of RMC and USOE to develop the Title I evaluation and reporting system were recognized, and the panel felt these efforts should not be wasted through mismanagement or lack of technical assistance. The suggestion was also made to include a system of data checks to ensure that meaningful data were collected and reported. Moreover, the panel was somewhat concerned that the tests might not be administered correctly in the classroom, an assumption that the reporting system takes for granted.

Some participants indicated that the underlying assumption for Model A, namely that a child is expected to maintain its relative status with respect to the national norm group from pretest to posttest without treatment, may be reasonable, but needs to be verified empirically.

Another problem involved labeling NCE gains. Since no empirical distribution of NCE gains exists, the panel felt that it was premature to state with assurance how large an "exemplary" NCE gain should be. Furthermore, the rule of thumb proposed by RMC for determining good projects, namely a gain of one-third of a standard deviation, needed empirical verification. The panel indicated that both issues could be resolved by a field test.

An important question directed to the U.S. Office of Education representatives concerned whether the testing of the no-treatment groups used in Models B or C would become a legitimate Title I expenditure if the system were mandated. Although no conclusions were reached, USOE indicated it would consider the problem.

Panel Recommendations

The recommendations made by the Policy Advisory Panel were directed to both RMC and the U.S. Office of Education. Immediate action is taking place on some issues, such as the clarification by RMC of any terms (e.g., "project") that might be misunderstood. However, most of the other recommendations will take time and much careful thought before any action can be initiated.

The panel recommended that USOE reconsider its views on the field test. It was felt that a controlled field test of all models would be important, not only to determine the technical assistance needed for model implementation, but also to verify underlying model assumptions and determine an NCE gain distribution. The panel felt that these and other findings would be important in facilitating nationwide implementation.

The participants reiterated a sense of urgency concerning a guidebook to the system. As indicated earlier, RMC is currently working on a monograph that details model selection and implementation procedures.

The panelists also felt that USOE should give more thought to its technical assistance plans. It was generally agreed that assistance given directly by USOE to the SEAs was preferred to that given by an intermediate organization.

The problem of conveying the meaning behind the reporting system, and the issue of public relations in general, was directed to both RMC and USOE. It was felt that both Congress and the educational community need to be informed about NCEs before the evaluation and reporting system can be implemented. At the present time, Congressional reports on educational program impact are expressed in average monthly gains. This suggests that a major reorientation effort may be necessary.

Although RMC is currently writing a computer program that will aid SEAs and LEAs in their data analyses for each model, the panel felt that USOE should consider developing machine-readable forms for data collection. It was believed that such forms, in addition to making the data-collection process more efficient, would also tend to reduce the human error (e.g., transcription, calculation) that would otherwise be transmitted through the system.

As a further aid to States with large numbers of districts, the panel recommended that the USOE offer sampling as an alternative to evaluations of all Title I projects and children. Appropriate sampling procedures would be developed and specified in sufficient detail so that the results would be both meaningful and statistically valid.

The panel was particularly interested in the exact form that the regulations concerning testing times would take. Since Model A requires testing at times when normative data were collected, the participants felt strongly that fall and spring testing should not be required because of the limited number of achievement tests with twice-a-year norms.

Several other recommendations were expressed. Among these were:

- USOE should determine and specify the exact minimum data requirements needed for Federal decision making.
- USOE personnel should consider what they will do with the data they receive, what they will present to Congress and the public, and what feedback they will provide to the States.
- USOE and NIE should coordinate similar Title I evaluation studies, so that the results can be synthesized for identification of future needs. Such a plan would eliminate duplication of data collection or analysis efforts.
- USOE should consider a longitudinal study of Title I, rather than an annual evaluation.

V. TECHNICAL ADVISORY PANEL MEETING

Description of the Meeting

During the first month of the contract and prior to conducting site visits, RMC convened a meeting of recognized experts in educational evaluation and research design. This meeting was held July 9, 10, and 11, 1975, in Palo Alto, California. The purpose of the conference was to review the evaluation models from all methodological viewpoints. This review included discussions of the strengths, weaknesses, and hazards encountered for each model.

A list of the participants is found in Table 3. The invited experts were selected to represent a wide range of educational evaluation areas. Their areas of expertise included: large-scale evaluation studies (Glass and Lord), criterion-referenced evaluation (Nitko), regression designs (Reichardt), control group designs (Lord), and testing and human abilities (Snow).

Technical Discussions

Highlights of the conference discussions, including comments on the system, each model, and the common metric, are presented below.

General Comments. Several panel members reflected on the problems of using quasi-experimental rather than true experimental designs. It was generally recognized that evaluators do not have complete control over all aspects of their Title I programs, such as designing the treatment, choosing testing times, assigning students randomly to treatment or no-treatment groups, and so on. Consequently, quasi-experimental designs may be the only valid method for evaluating these programs. However, several participants felt that strict experimental designs should not be overlooked but set as an "ideal" goal because only well controlled experiments can partial out the very subtle biases found in applied research. All panelists agreed that an evaluator should try to use the best evaluation techniques that he can with the situation at his disposal

One concern shared by most of the panel members was that Title I may benefit the students who are not participating in the special programs because the teacher-student ratio is improved when the Title I children are pulled out of the regular classroom to receive the treatment. This situation would invalidate Model C and make the finding of equivalent control groups for Model B even more difficult.

Table 3

TECHNICAL ADVISORY PANEL MEETING
July 9, 10, 11, 1975

TECHNICAL ADVISORY PANEL

Gene Glass	University of Colorado
Frederic Lord	Educational Testing Service
Anthony Nitko	University of Pittsburgh
Charles Reichardt (1)	Northwestern University
Richard Snow	Stanford University

U.S. OFFICE OF EDUCATION

Janice Anderson	Office of Planning, Budgeting, and Evaluation
William Lobosco	Office of Planning, Budgeting, and Evaluation
George Mayeske	Office of Planning, Budgeting, and Evaluation

RMC PROJECT STAFF

Alexander Cannara	Richard Johnson
Alice Chiang	Sonia Johnson
Barbara Fagan	Larry Rosen
Nona Gamel	G. Kasten Tallmadge
Donald Horst	Christine Wood

(1) Attended for Donald Campbell.

Several issues related to criterion-referenced tests were discussed at great length. One concerned the implications of scoring CRTs by items rather than by objectives. Since item scoring produced the larger score distribution, the panel felt this method would yield data more amenable to calculations of means and standard deviations.

A more critical discussion focused on the method of aggregating CRT and NRT scores. Initially, some panel participants were not sure whether the sample or the population standard deviation should be used to adjust the gains obtained using CRTs. They later agreed that, insofar as a CRT accurately measured its particular content area (i.e., a reading CRT actually measures reading skills), the aggregation procedures based on the CRT population standard deviation were reasonable, and no major implementation problems were foreseen.

An interesting suggestion, raised by a few panelists, concerned the possibility of using data from previous years to establish a baseline with which succeeding years of Title I could be compared. To the extent that such data can be converted into local norms, this recommendation has been incorporated into the system.

System comments. The panel spent a considerable amount of time discussing the merits of affective data--particularly whether teacher judgments should be collected as a part of the evaluation and reporting system. A few panel members felt strongly that professional judgments should be used to supplement the data from standardized tests to form a more complete picture of Title I impact. Reactions to this position were several: since teacher judgments are difficult to quantify and no valid collection procedures for these kinds of data have yet been devised, they should not be included as part of the system; anecdotes and narratives, which were often included in States' evaluation reports, are not useful nor desired; some narratives might be useful to supplement the achievement data, but they would not have to be collected from each project; and affective data are recognized as being potentially useful but would not add significantly to the current efforts to develop the evaluation and reporting system.

System comments: Model A. The appropriateness of a test publisher's national norms was discussed at length. Since a norming sample typically includes only those schools willing to cooperate with the test publisher, the panel felt that the norms are not likely to be representative of the entire student population. The panel members also felt that publishers' norm groups might differ slightly from the Anchor Test Study norm group (Loret et al., 1974) since the latter included private and church-related schools, while in some instances, the former did not. Upon further examination of norm-group composition, this problem was judged to be minor.

The question of when to administer a standardized test also generated lengthy discussions. A few panelists felt that the requirement to test at times that correspond to empirical normative data points was much too severe. They argued that, since the magnitude of the interpolation errors is unknown, a hard line on the testing issue is unnecessary. Others agreed that this effort may be a form of misplaced precision but argued that these errors were ones that could be controlled. Since the system must contain many sources of error, some known and some unknown at the present time, the ability to identify an error source and then reduce it was considered a positive step.

To facilitate implementation of Model A, the suggestion was made to allow extrapolations of six weeks on both sides of the publisher's data points. It was tentatively proposed that double-entry norms tables be developed that would provide the extrapolation by means of a nomograph. This issue was discussed with several of the major test publishers during the course of the year, but little progress has been made in actually constructing the tables.

The panel also observed that the equating of different test levels and forms was less than perfect and probably not adequate for norm-referenced evaluations. They agreed that pretesting and posttesting with the same test level and form was preferable, whenever possible.

The implications of selecting students on the basis of low pretest scores was discussed at great length. This procedure artificially inflates the gain from pretest to posttest due to statistical regression to the mean. Several discussions followed over whether the regression effect would occur even with perfectly reliable tests and whether the treatment group would regress toward the mean of the national distribution or toward the mean of the Title I population. In any case, all agreed that the treatment children should be pretested only after they had been selected.

Other issues concerning the norm-referenced model were also discussed. Some of these were:

- The bottom tail of a national norming distribution may not be an appropriate comparison for Title I children since the two groups are likely to differ in racial and socioeconomic composition and other educationally relevant ways.

- Recent publishers' norms may include Title I children, thus contaminating the "control" group.

- There is a practice effect of pretest on posttest, which may raise the Title I scores as compared to the norm sample.

- Extreme groups always show some regression toward the mean even with "true" scores; thus, gain estimates may be somewhat inflated.

System comments: Model B. An important issue was discussed with respect to the control group model, namely, whether randomly assigned control groups, or even "random in effect" groups, are likely to be available. Several panel members felt that some of the "borderline" schools that do not receive Title I funds could be used effectively as controls for those "borderline" schools that receive Title I funds. However, the point was made that these non-Title I schools, especially in the larger States, are extremely likely to receive State compensatory funds in lieu of Title I money. Thus, it is likely that low-achieving children in a "borderline," non-Title I school will be receiving services from a State-funded compensatory program and could not, therefore, provide a meaningful no-treatment expectation.

The participants also discussed methods for adjusting posttest scores to compensate for pretest differences resulting from the use of nonequivalent control groups. Some panel members felt that adequate adjustments to account for all pre-existing differences between the groups, which requires a complete list of the relevant covariates, were simply not available. Other panelists felt that, even with these imperfect adjustments, Model B yielded the best no-treatment expectation of any of the models.

System comments: Model C. Discussion of the special regression model centered around data interpretation. It was suggested that any curvilinearity of regression might falsify conclusions drawn from a Model C evaluation. It was pointed out, however, that curvilinear regression would not occur with well constructed tests unless test ceilings or floors were encountered. The panel also worried that the Regression Projection Model might be subject to intolerably large standard errors, even when the assumptions of the model were met. This fear no longer seems justified.

System comments: NCEs. The normality assumptions used to obtain the NCEs were questioned by some panel participants. They felt that these assumptions might not hold for all test publishers' data, since the original raw-score distributions are "massaged" into a normal distribution. To determine how the test score distributions were actually generated, RMC staff members queried several test publishers over the year. Their findings suggested that, at least for the major test publishers, the normality assumptions underlying the Normal Curve Equivalent were valid.

A problem in interpreting the system's results was also discussed. Several panel members pointed out that NCE gains could not be directly

correlated with success in school, the ability to read a newspaper, the ability to read a page of fourth-grade text, and so on. They felt that the issue of what the gains mean in terms of specific progress for the children involved should be considered in more detail.

The RMC project staff found the Technical Advisory Panel's comments extremely helpful. Although it is not clear that any changes to the system sprang directly from the meeting (the use of local norms with Model A1 may be an exception), several technical issues were brought more sharply into focus. Additional thinking was stimulated by the discussions and insights were gained that shaped subsequent modifications to the system. Of most value, perhaps, was the opportunity to work with some of the leading experts in the field and to have them confirm the soundness of the thinking that had gone into the system.

VI. TECHNICAL DEVELOPMENTS

During Phase I many technical issues were confronted. Some of those that resulted in modifications to the system are discussed in this chapter. Among these are the use of local norms and non-normed tests with Model A, and the addition of adjustment techniques to compensate for nonequivalent control groups in Model B. The issues that will be investigated during Phase II are described in Chapter VII.

All of the technical developments described here resulted from efforts to make the system more attractive to its users. It was believed that a flexible system that can accommodate many diverse evaluation situations is more useful than one that can be used only in a limited number of settings. The SEA and LEA personnel who were visited during the year provided valuable feedback concerning additional features of the system that they considered both highly desirable and important for successful nationwide implementation of the system.

One of the criticisms directed at Model A has been the use of a national norming sample to obtain a no-treatment expectation for Title I children. The objection is frequently made that the norm sample does not constitute an appropriate "control" for disadvantaged children. An ideal evaluation would involve two identical groups of children: one group would receive the special treatment while the other would not. At the end of the project, the no-treatment group's performance would be compared with that of the treatment group to arrive at an estimate of the project's impact. However, since it is almost impossible to duplicate the conditions of the ideal evaluation in actual educational settings, it is generally recognized that the more similar the children in the norming sample are to the children in the treatment group, the more accurate the no-treatment expectation will be.

The fact that some subgroups may increase their status relative to the national norms while others may go down invalidates the norm-referenced model since its underlying assumption is that, without special treament, students will maintain their status with respect to a norm group. Thus, a local norm group, which can be constructed to contain children that are highly similar to the treatment children on whatever educationally relevant variables are selected, may provide a more accurate no-treatment expectation than the national norm group.

The locally normed test does not have to have national norms in order to provide a no-treatment expectation. However, to convert the local gains to NCE gains for aggregation purposes, a nationally normed test must also be administered. The procedures for obtaining NCE gains are as follows: (a) subtract the treatment group's no-treatment pretest score from the observed posttest score, (b) divide this gain by the standard deviation of the national norm group on the nationally normed test, and (c) multiply this quantity by 21.06 to convert it to NCEs.

The forms and instructions are currently designed to accommodate only locally normed tests that also have national norms. If the locally normed test lacks national norms, the norm-referenced model can still be implemented, but the necessary worksheets and reporting forms will have to be improvised.

Feedback from SEAs and LEAs over the year also suggested that a modification to Model A, considered by many to be the most popular model, was needed to accommodate non-normed tests. To this end, Model A2 was developed. The forms and instructions, which have been designed to accommodate non-normed tests, provide procedures for deriving appropriate no-treatment expectations and expressing the project gains in NCEs. A detailed description of this model can be found in Chapter II (pages 18 and 19).

To extend Model B's applicability, provisions were made to allow the use of control groups that might not be strictly comparable. Ideally, the treatment and control groups should come from the same population and thus, should not differ significantly on any educationally relevant variables. In actual educational settings, however, intact groups that are designated as treatment or control may differ markedly in terms of socioeconomic status or racial composition. Any pretest differences found between these two groups are probably real, since assignment to the treatment and control groups cannot be considered random or random in effect.

To compensate for these real pretest differences, which are not due to random events, the principal-axis or standardized gain score method of adjustment (Kenny, 1975) is used. The analysis of covariance method, which was formerly the only adjustment procedure recommended, is now only used when the treatment and control groups can be regarded as random samples from a single population.

The forms and instructions present step-by-step procedures to determine which, if any, of the two adjustment methods are appropriate, and to perform the necessary calculations for the selected method.

In Model C, the Technical Advisory Panel's discussions of the conditions that might falsify the special regression model's conclusions generated much thought on the part of the RMC staff. Although none of these discussions resulted in any modifications to the model, some of the issues such as the effects of test ceilings and floors, the effects of differing sample sizes for the treatment and comparison groups, and the stability of the comparison group's regression line, will continue to be examined during Phase II.

VII. PHASE II ACTIVITIES

In their present form, the evaluation models are both operationally and technically sound. During the developmental stage and the subsequent site visits, however, several questions have arisen concerning technical and implementation issues. Both USOE and RMC consider these issues sufficiently important to warrant further investigation. A brief description of these Phase II activities is presented below:

1. Norming and score conversion practices of major test publishers will be investigated and a working paper will be written describing these practices and their impact on the evaluation models.

   The paper will briefly describe the norming practices used by each of the major test publishers and the sources of inaccuracy that exist in normative data. The scaling techniques that are used to generate the derived scores from the normative data will also be reviewed.

2. A computer program for complete implementation of all models will be written and documented.

   The program will be written in Fortran, a language commonly available on a wide variety of machines. It will be documented in both its text and a brief manual. The program, which will follow the structure of the reporting system as described in Chapter II, is designed to be flexible in terms of the following variables:

   - Types of scores: although the program is designed to accept percentile scores or NCEs as input, it will also accommodate raw or standard scores if the user provides the appropriate conversion tables.

   - Input devices: the program will accept any input device, such as cards, tapes, or disks.

   - Error detection: the program will provide appropriate error messages to the user at all important points and at all levels of output reporting.

3. Real and simulated data will be used to study the properties of the models.

   Both actual and fictional data will be used to examine such real-world effects as test floors and ceilings, small sample sizes, inappropriate selection procedures for treatment children, and so on.

4. A set of procedures will be formulated to use in compiling data from the evaluation reports.

   Compilation at the Federal level requires two levels of analysis. On a general level, forms and instructions will be provided to aggregate the six types of information requested by the system: student participation, parent advisory councils, personnel, training, cost, and cognitive impact data. On a more detailed level, procedures will be provided for interpreting these results, the impact data in particular.

5. A "mock-up" State report will be produced using real data and the analysis procedures recommended by the evaluation models.

   A sample State report, which will contain actual data from several projects, will be written to illustrate how the data are analyzed using one of the evaluation models and how the results from each model are aggregated to produce the final report. Models A and C are the most likely models for inclusion in the report.

6. The State Title I Annual Evaluation Reports for FY 75 will be described in "critical summary" narratives.

   Content summaries of the State reports will be prepared to provide a baseline with which future reports, conforming to the Title I evaluation and reporting system, will be compared. Each report's contents will be described with respect to the types of information reported in the following areas: target population, expenditures, impact data, needs assessment, parent advisory councils, and other system-relevant data.

7. The costs incurred as a result of system implementation will be analyzed.

   An attempt will be made to specify potential sources of additional costs. Among these might be: personnel, computer services, record-keeping procedures, and testing no-treatment children.

REFERENCES

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Campbell, D. T., & Erlebacher, A. E. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), Disadvantaged child. Vol. 3. Compensatory education: A national debate. New York: Brunner/Mazel, 1970.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Chiang, A., & Rosen, L. D. The Normal Curve Equivalent: A metric to measure instructional project impact. Submitted for publication.

Gamel, N. N., Tallmadge, G. K., Wood, C. T., & Binkley, J. L. State ESEA Title I reports: Review and analysis of past reports, and development of a model reporting system and format. Mountain View, Ca.: RMC Research Corporation, October 1975. (RMC Report No. UR-294)

Glass, G. V. Data analysis of the 1968-1969 survey of compensatory education (Title I). Final Report. Boulder, Co.: University of Colorado/ Laboratory of Educational Research, August 1970.

Horst, D. P., Tallmadge, G. K., & Wood, C. T. A practical guide for measuring project impact on student achievement. Washington, D.C.: U.S. Government Printing Office, 1975. (Stock No. 017-080-01460) Also published as Measuring achievement gains in educational projects. Los Altos, Ca.: RMC Research Corporation, October 1974. (RMC Report No. UR-243)

Kenny, D. A. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. Psychological Bulletin, 1975, 82, 345-362.

Levine, R. S., & Angoff, W. H. The effects of practice and growth on scores on the Scholastic Aptitude Test. Princeton, N.J.: Educational Testing Service, February 1958. (R and DR No. 58-6/SR-58-6)

Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. A. Anchor test study: Equivalence and norms tables for selected reading achievement tests (grades 4, 5, 6). Washington, D.C.: U.S. Government Printing Office, 1974. (Office of Education Report 74-305)

REFERENCES (continued)

Tallmadge, G. K. Cautions to evaluators. Paper presented at the U.S. Office of Education's Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Reston, Virginia, 1976.

Tallmadge, G. K., & Horst, D. P. A procedural guide for validating achievement gains in educational projects. Mountain View, Ca.: RMC Research Corporation, December 1975. (RMC Report No. UR-240) Also published under the same title in Washington, D.C.: U.S. Government Printing Office, 1976. (Stock No. 017-080-01516)

Thorndike, R. L. Regression fallacies in the matched groups experiment. Psychometrika, 1942, 7, 85-102.

U.S. Office of Education. Education of the disadvantaged: An evaluation report on Title I Elementary and Secondary Education Act of 1965, fiscal year 1968. Washington, D.C.: U.S. Government Printing Office, 1968.

U.S. Office of Education. Joint federal/state task force on evaluation: An overview (Belmont System), February 1971.

U.S. Department of Health, Education, and Welfare/Office of Education. Statistical report, fiscal year 1968: A report on the third year of Title I ESEA. Washington, D.C.: U.S. Government Printing Office, January 1970.

Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin, 1976, 83 (2), 213-217.

Wargo, M. J., Tallmadge, G. K., Michaels, D. D., Lipe, D., & Morris, S. J. ESEA Title I: A reanalysis and synthesis of evaluation data from fiscal year 1965 through 1970. Palo Alto, Ca.: American Institutes for Research, March 1972. (AIR-27400-3/72-FR)